27-03-2020

PREDICTIVE ANALYSIS PROJECT

CUSTOMER CHURN FOR TELECOM COMPANIES

PRESENTED BY: SHILPA GIRIDHAR

CONTENTS

DESCRIPTION

Problem Statement

Customer Churn is a burning problem for Telecom companies. In this project, we simulate one such case of customer churn where we work on a data of postpaid customers with a contract. The data has information about the customer usage behaviour, contract details and the payment details. The data also indicates which were the customers who cancelled their service. 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.

Variables

Churn	1 if customer cancelled service, 0 if not
AccountWeeks	number of weeks customer has had active account
ContractRenewal	1 if customer recently renewed contract, 0 if not
DataPlan	1 if customer has data plan, 0 if not
DataUsage	gigabytes of monthly data usage
CustServCalls	number of calls into customer service
DayMins	average daytime minutes per month
DayCalls	average number of daytime calls
MonthlyCharge	average monthly bill
OverageFee	largest overage fee in last 12 months
RoamMins	average number of roaming minutes

Requirements

Perform the following :

1. **EDA (16 Marks)**
   - How does the data looks like, Univariate and bivariate analysis. Plots and charts which illustrate the relationships between variables (4 Marks)
   - Look out for outliers and missing values (4 Marks)
   - Check for multicollinearity & treat it (4 Marks)
   - Summarize the insights you get from EDA (4 Marks)
2. **Build Models and compare them to get to the best one (39 Marks)**
   - Applying and interpreting Logistic Regression (8 Marks)
   - Applying and interpreting KNN (8 Marks)
   - Applying and interpreting Naive Bayes (8 Marks) (is it applicable here? comment and if it is not applicable, how can you build an NB model in this case?)
   - Model Comparison using Model Performance metrics & Interpretation (15 Marks) - Confusion matrix interpretation for all models, Confusion matrix interpretation for all models, Remarks on Model validation exercise <Which model performed the best>
3. **Actionable Insights (5 marks)**
   - Interpretation & Recommendations from the best model (5 Marks)

BUSINESS OBJECTIVE

When a company's revenue is based on recurring monthly or annual subscription with customer, every customer who leaves puts a dent in your cash flow. So, understanding why customers cancel their subscriptions is important for companies to address the issue. Building a predictive churn model helps companies to take some action to retain such customers. A predictive churn model extrapolates on existing data (the number of customers who left your service during a given time period) to show future potential churn rates.

The dataset contains customer-level information for a span of 1 Week to 4.6 years (243 weeks).

To guide the analysis, we are going to try and answer the following questions about my customer segments:

- Does the number of calls made to Customer Service indicate the individuals who are more likely to churn?
- Do individuals with Data Plan and Data Usage more like to churn more than those without a Data Plan?
- Does the number of Calls made per month or the Roaming Calls made per month significant in classifying individuals pattern?
- Does the Monthly Charges or the Overage Fees indicate the individuals who are more likely to churn?

Variables	
Churn	1 if customer cancelled service, 0 if not
AccountWeeks	number of weeks customer has had active account
ContractRenewal	1 if customer recently renewed contract, 0 if not
DataPlan	1 if customer has data plan, 0 if not
DataUsage	gigabytes of monthly data usage
CustServCalls	number of calls into customer service
DayMins	average daytime minutes per month
DayCalls	average number of daytime calls
MonthlyCharge	average monthly bill
OverageFee	largest overage fee in last 12 months
RoamMins	average number of roaming minutes

DATA EXPLORATION

- The data shows that there are 3333 observations and 11 variables
- Performed the str and summary function
- Converted “Churn”, “ContractRenewal”, “DataPlan”, “Customer Service Calls” , and “AccountWeeks“ into factor variables
- Check for any missing values – No missing values in data set
- Churn column is the target variable

```
 Churn      AccountWeeks   ContractRenewal DataPlan   DataUsage       CustServCalls
 0:2850   Min.    :  1.0   0: 323          0:2411   Min.    :0.0000   1       :1181
 1: 483   1st Qu.: 74.0   1:3010          1: 922   1st Qu.:0.0000   2       : 759
          Median :101.0                            Median :0.0000   0       : 697
          Mean   :101.1                            Mean   :0.8165   3       : 429
          3rd Qu.:127.0                            3rd Qu.:1.7800   4       : 166
          Max.   :243.0                            Max.   :5.4000   5       :  66
                                                                    (Other):  35
    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
```

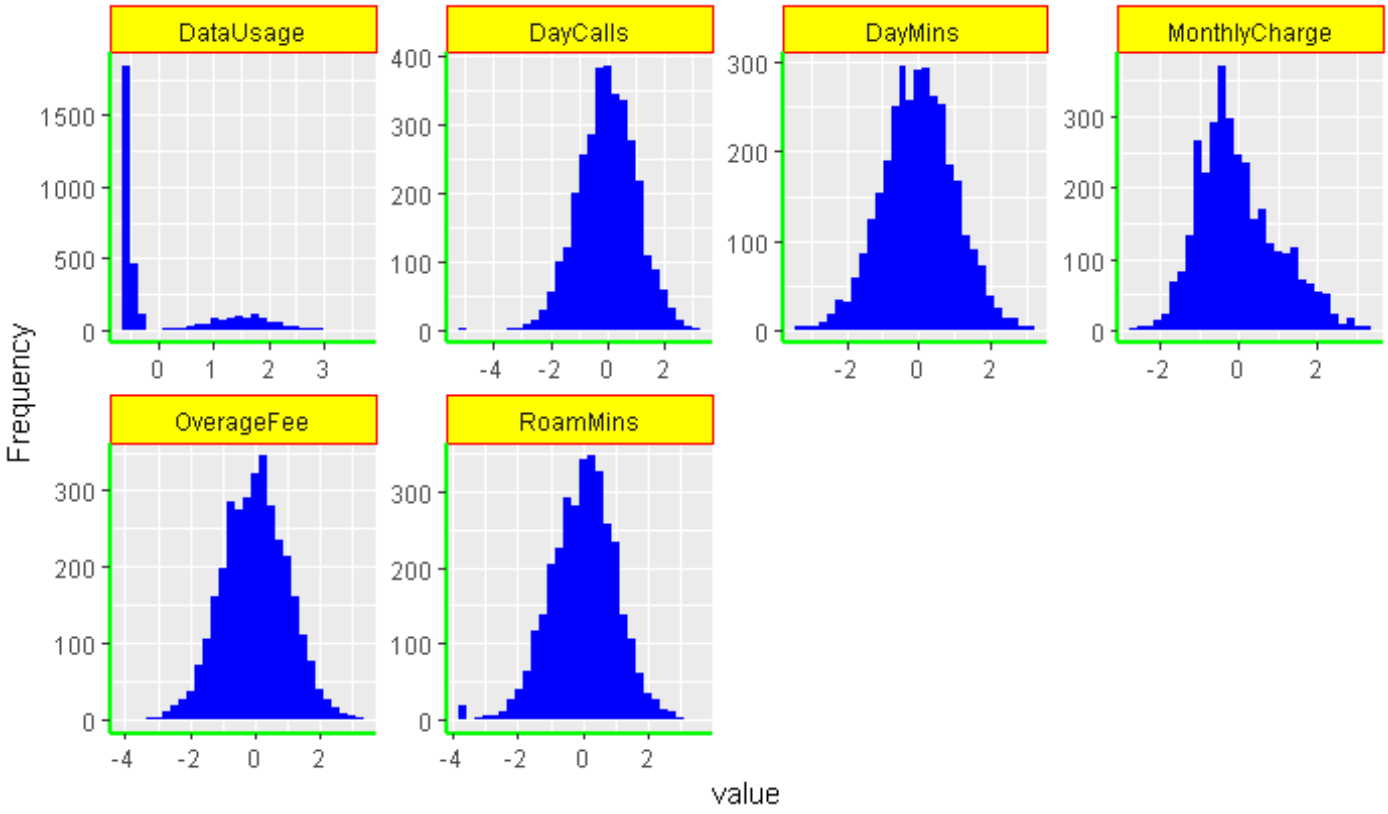
- “AccountWeeks” has too many levels. We group in order to reduce the number of levels.

Account weeks has been categorized into six months groups: “0–6 Month”, “6–12 Month”, “12–18 Months”, “18–24 Month”, “24–30 Month”, “30–36 Month”, “36–42 Month”, “42–48 Month”, “48–54 Month”, “54–60 Month”

- The output above confirms that the numerical variables have different units and scales, for example, ‘Data Usage’ in gigabytes and ‘Monthly Charge’ in rupees. These differences can unduly influence the model and, therefore, we need to scale or transform them. We need to normalize data using Standardization technique in which all the variables are centred around zero and have roughly unit variance. The ‘preprocess’ function in Caret library is used to transform dataset.

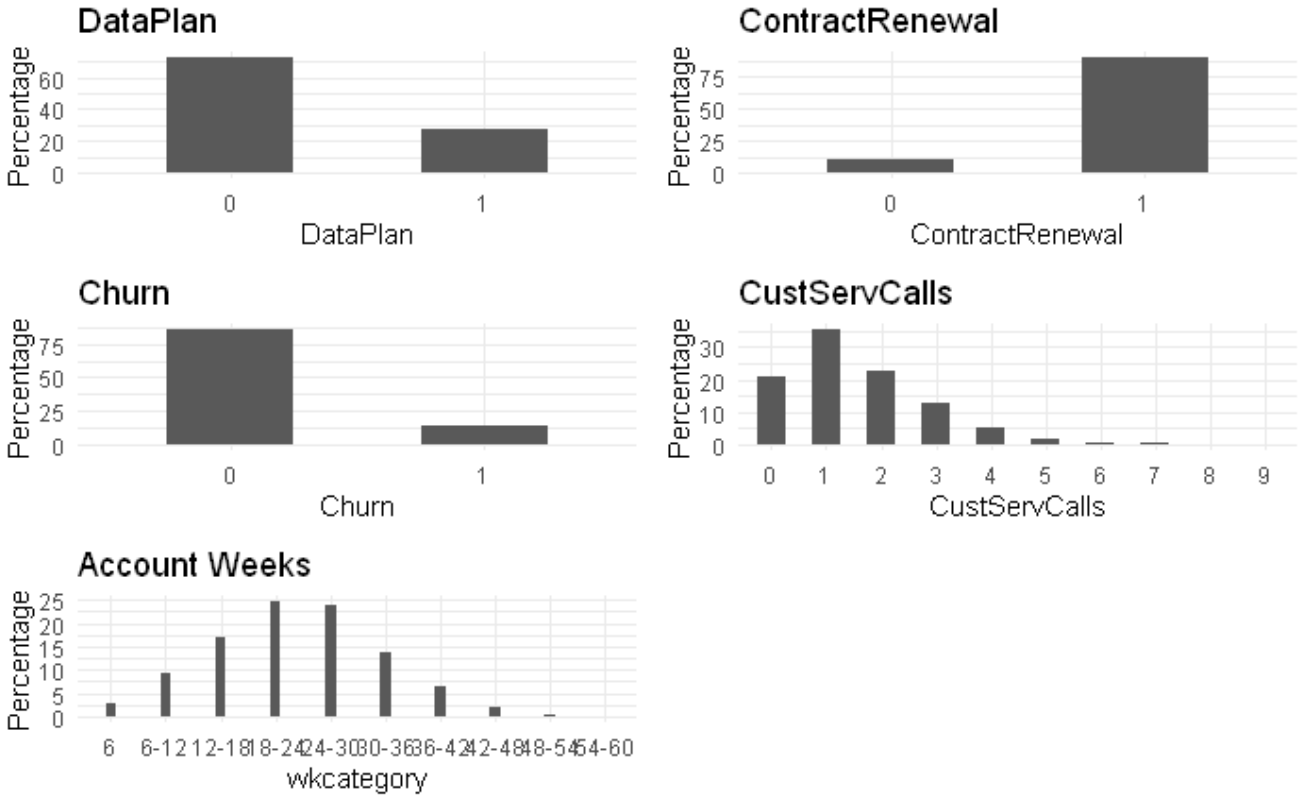
```
 Churn    ContractRenewal DataPlan   DataUsage       CustServCalls     DayMins
 0:2850   0: 323          0:2411   Min.    :-0.6415   1       :1181   Min.    :-3.300601
 1: 483   1:3010          1: 922   1st Qu.:-0.6415   2       : 759   1st Qu.:-0.662325
                                   Median :-0.6415   0       : 697   Median :-0.006887
                                   Mean   : 0.0000   3       : 429   Mean   : 0.000000
                                   3rd Qu.: 0.7571   4       : 166   3rd Qu.: 0.672419
                                   Max.   : 3.6015   5       :  66   Max.   : 3.139950
                                                     (Other):  35
    DayCalls        MonthlyCharge       OverageFee        RoamMins          wkcategory
 Min.   :-5.00450   Min.   :-2.5755   Min.   :-3.9640   Min.   :-3.66686   18-24  :822
 1st Qu.:-0.66947   1st Qu.:-0.6882   1st Qu.:-0.6789   1st Qu.:-0.62228   24-30  :797
 Median : 0.02812   Median :-0.1708   Median : 0.0073   Median : 0.02246   12-18  :561
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   30-36  :460
 3rd Qu.: 0.67588   3rd Qu.: 0.6024   3rd Qu.: 0.6777   3rd Qu.: 0.66720   6-12   :305
 Max.   : 3.21711   Max.   : 3.3480   Max.   : 3.2096   Max.   : 3.49687   36-42  :211
                                                                           (other):177
```

- Use Histogram plot to understand continous variables

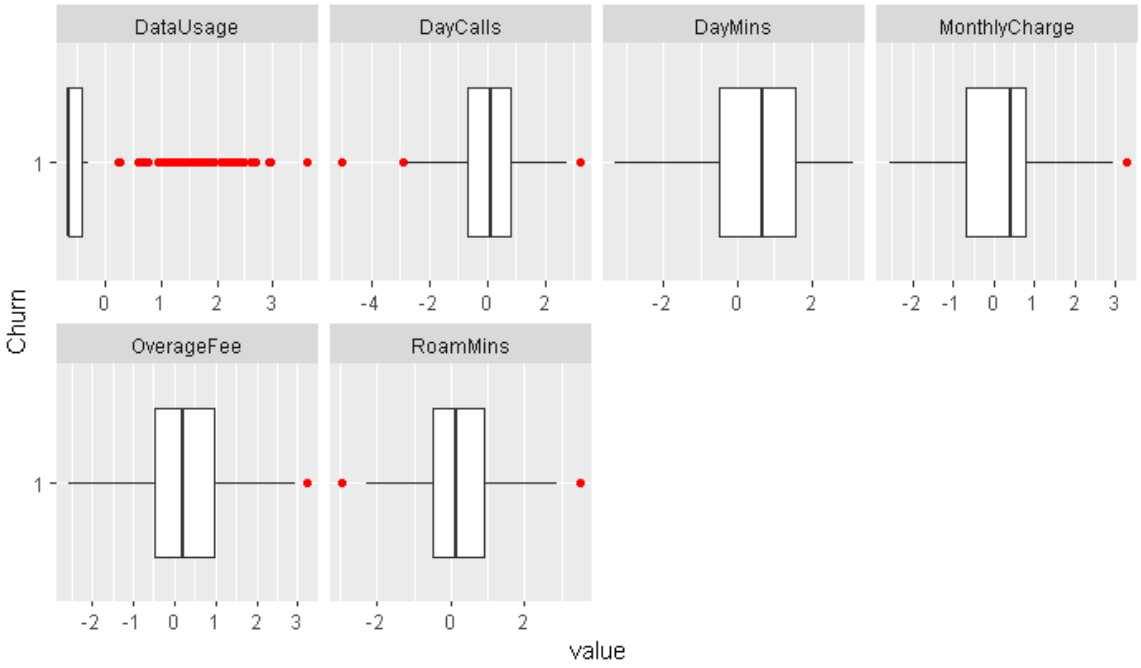


- Use Bar plot to understand categorical variables
  Churn = 1/ Contract Renewal = 0 – are the customers who have cancelled the service
  Churn = 0/ Contract Renewal = 1 – are the customers who have not cancelled the service
  Data Plan = 1 – are the customers who have a Data Plan
  Data Plan = 0 – are the customers who do not have Data Plan

  Majority of the customers have not opted for the data plan. Majority of the customers have recently renewed the service and not cancelled the service is much higher than the customers who have cancelled the service (Churners). Dataset in imbalanced.

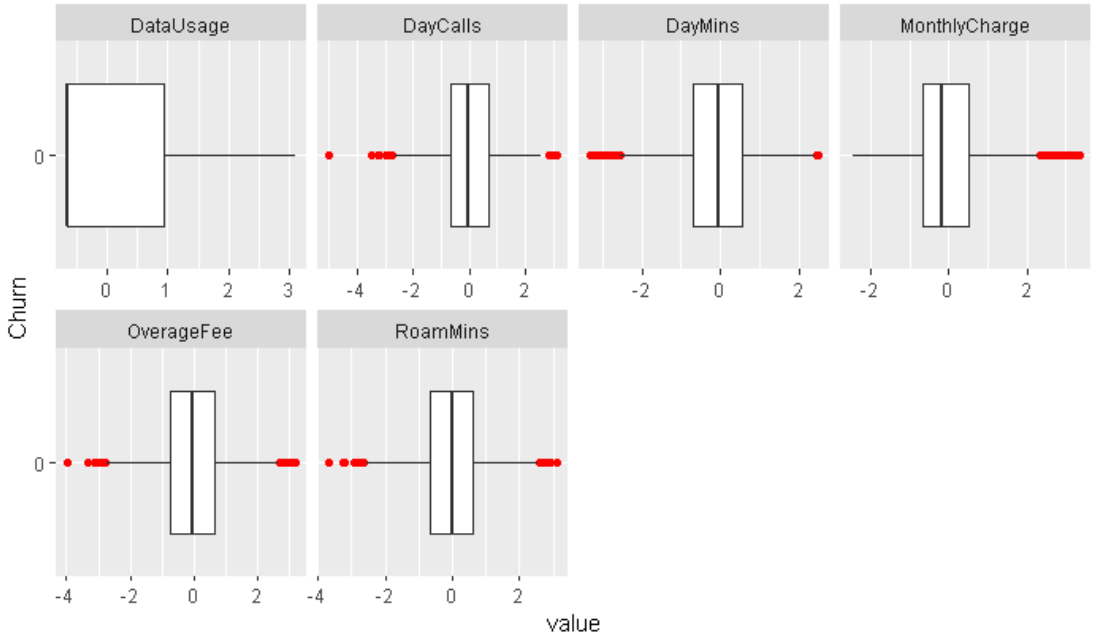


- Using boxplot grouped by “Churn” column data
  We further filter the main data set to get subset “ds1” that has data of all the customers who have churned, that is with Churn ==1.Data usage have outliers for category of customers who have cancelled their service that we should investigate. Monthly charge has the highest variability in scores and is potentially left-skewed. Potential variables that may affect the churn are the number of calls, monthly charges, overage fee or roaming minutes. Data usage implication seems negligible.



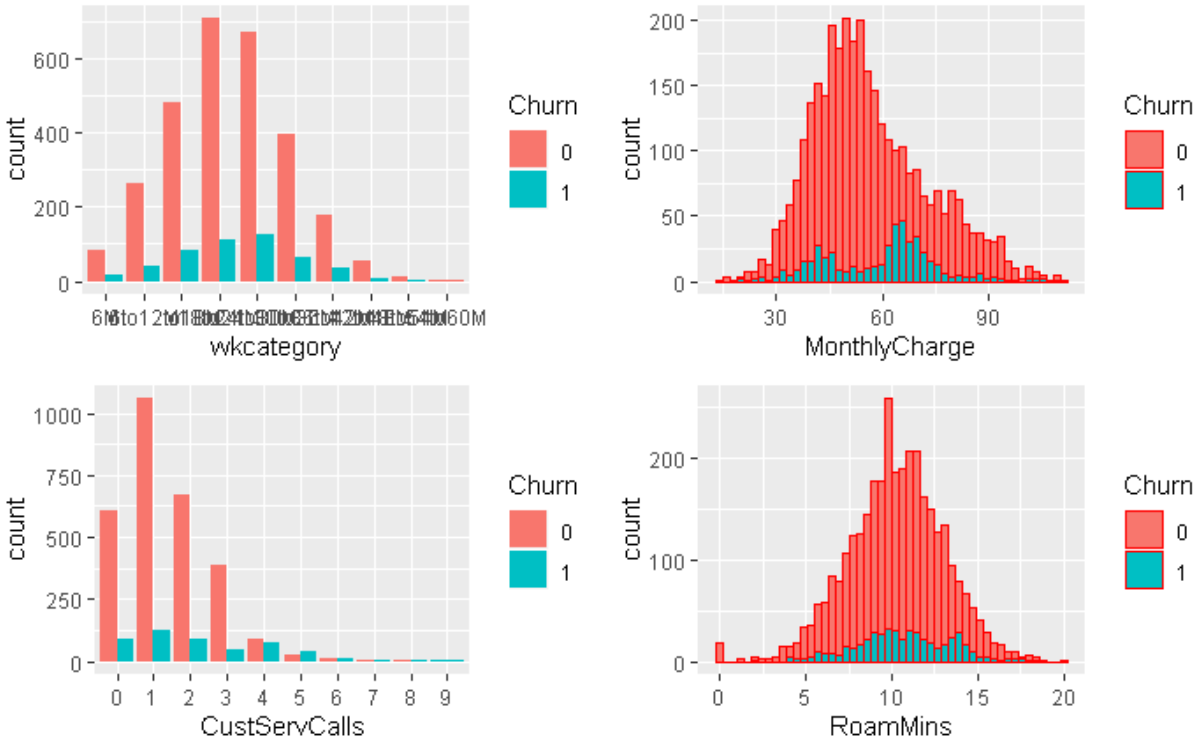
Box plot for Churners

All other variables have many outliers in the Non-Churner category (Churn = 0)

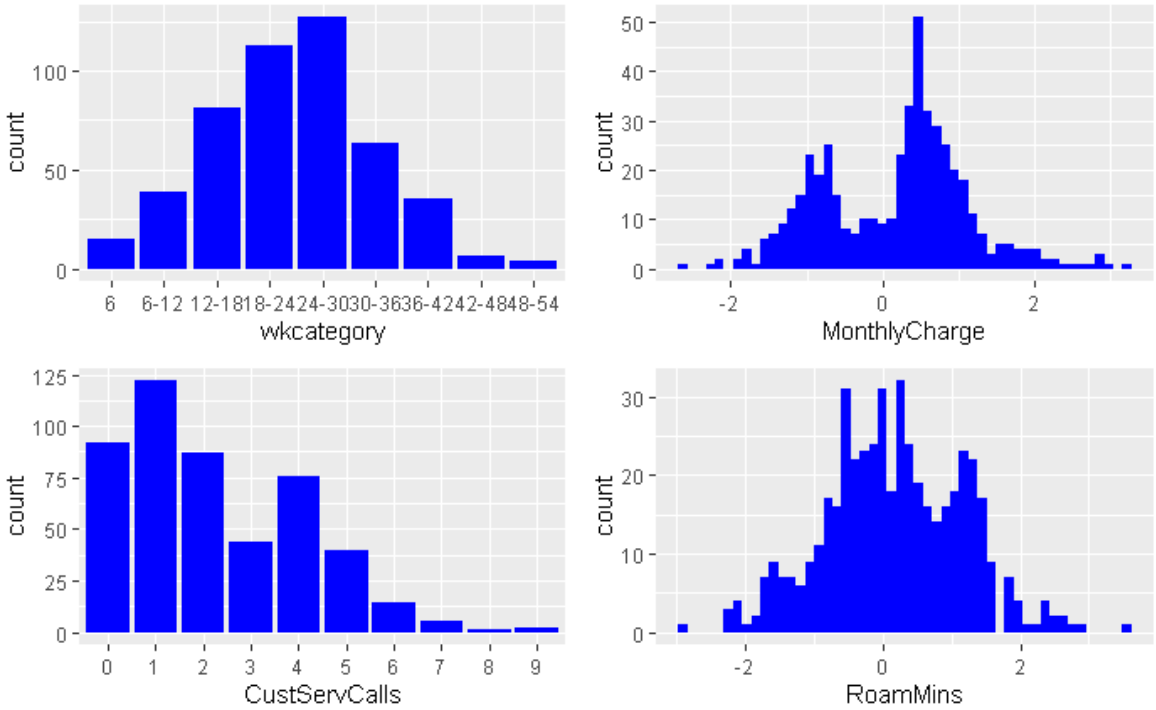


Box plot for Non-Churners

- Bivariate analysis

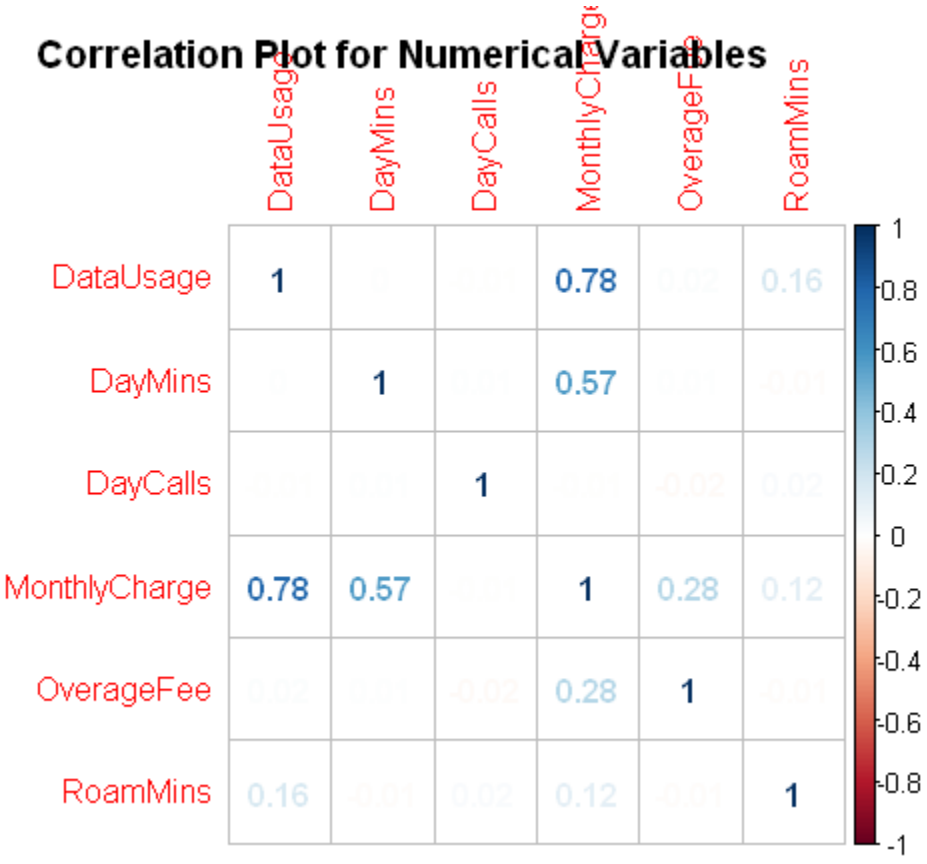


Zooming in on the Churn==1 customers --> Maximum number of customers who have churned lies in the 12 months to 36 months category



- Used the library(corrplot) to plot correlation and check for high correlation between numerical variables

  We observe that DayMins, Data Usage are highly correlated with and Monthly Charges.



MODEL EVALUATION

1. Random Forest
2. Logistic Regression
3. K- Nearest Neighbors
4. Naïve Bayes

APPLYING AND INTERPRETING LOGISTIC REGRESSION

- To make sure that we're not overfitting our model, we'll split the main data set into 70% of the data to be our training set, and 30% to be our test set. We'll train the model on the training set, and then test out its performance on the test set. To create the split we'll use the Caret package.

- Train and Test data set gives the following observations

```
> table(trainset$Churn)

   0    1
1995  339
> prop.table(table(trainset$Churn))

        0         1
0.8547558 0.1452442
```

- Applying logistic regression on the data set gives the following output:

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4081  -0.4689  -0.3301  -0.1949   3.2443

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.15086    0.45591   0.331   0.7407
ContractRenewal1 -2.11851    0.17373 -12.195  < 2e-16 ***
DataPlan1        -2.18663    0.68436  -3.195   0.0014 **
DataUsage         4.34536    3.05907   1.420   0.1555
CustServCalls1   -0.17322    0.19460  -0.890   0.3734
CustServCalls2    0.06095    0.21266   0.287   0.7744
CustServCalls3   -0.43430    0.27173  -1.598   0.1100
CustServCalls4    1.95228    0.26047   7.495 6.62e-14 ***
CustServCalls5    3.28797    0.38180   8.612  < 2e-16 ***
CustServCalls6    3.83381    0.55134   6.954 3.56e-12 ***
CustServCalls7    3.41069    0.76251   4.473 7.71e-06 ***
CustServCalls8  -10.05276  535.41132  -0.019   0.9850
CustServCalls9   16.13226  535.41127   0.030   0.9760
DayMins           3.49830    2.20654   1.585   0.1129
DayCalls         -0.02282    0.06915  -0.330   0.7414
MonthlyCharge    -4.94126    3.90787  -1.264   0.2061
OverageFee        1.64845    1.03083   1.599   0.1098
RoamMins          0.13668    0.07709   1.773   0.0762 .
wkcategory6-12   -0.29923    0.45267  -0.661   0.5086
```

```
wkcategory12-18   -0.05283    0.41409  -0.128   0.8985
wkcategory18-24   -0.13103    0.40332  -0.325   0.7453
wkcategory24-30    0.01402    0.40140   0.035   0.9721
wkcategory30-36   -0.17477    0.42189  -0.414   0.6787
wkcategory36-42    0.01663    0.46257   0.036   0.9713
wkcategory42-48   -0.91705    0.79236  -1.157   0.2471
wkcategory48-54    0.99612    0.98695   1.009   0.3128
wkcategory54-60   -9.69250  535.41136  -0.018   0.9856
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1934.3  on 2333  degrees of freedom
Residual deviance: 1437.2  on 2307  degrees of freedom
AIC: 1491.2

Number of Fisher Scoring iterations: 12
```

- Roaming Minutes and Customer Service Calls category 4,5,6,7 are significant variables for predicting Customer Churn, as is evident from the significance code '***', printed next to the p-value of the variable.
- The p-value, shown under the column, *Pr(>|t|)*, is less than the significance value of 0.05, which also suggests that there are statistically significant relationships between the variables
- While Contract Renewal variable has a negative effect, indicating those customers who have renewed the contract or who have Data plan are unlikely to churn. However the coefficient of Contract Renewal is significant.
- AIC value of 1491. Lets build other models and compare the AIC value.

- Multicollinearity in logistic regression is equally important as other types of regression. Building model2 by excluding variables that highly correlated increases AIC value. DayMins, MonthlyCharges Overage Fee turn out to be significant variables.

```
Call:
glm(formula = Churn ~ ContractRenewal + DayMins + MonthlyCharge +
    OverageFee + RoamMins, family = binomial, data = trainset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5788  -0.5504  -0.4250  -0.2966   2.9817

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.42247    0.14339  -2.946  0.00322 **
ContractRenewal1 -1.81879    0.15993 -11.373  < 2e-16 ***
DayMins           0.83604    0.08597   9.724  < 2e-16 ***
MonthlyCharge    -0.44005    0.09057  -4.858 1.18e-06 ***
OverageFee        0.38552    0.06930   5.563 2.65e-08 ***
RoamMins          0.20396    0.06463   3.156  0.00160 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1934.3  on 2333  degrees of freedom
Residual deviance: 1673.3  on 2328  degrees of freedom
AIC: 1685.3
```

```
Number of Fisher Scoring iterations: 5
```

- Evaluate the performance of model 1 using Confusion Matrix

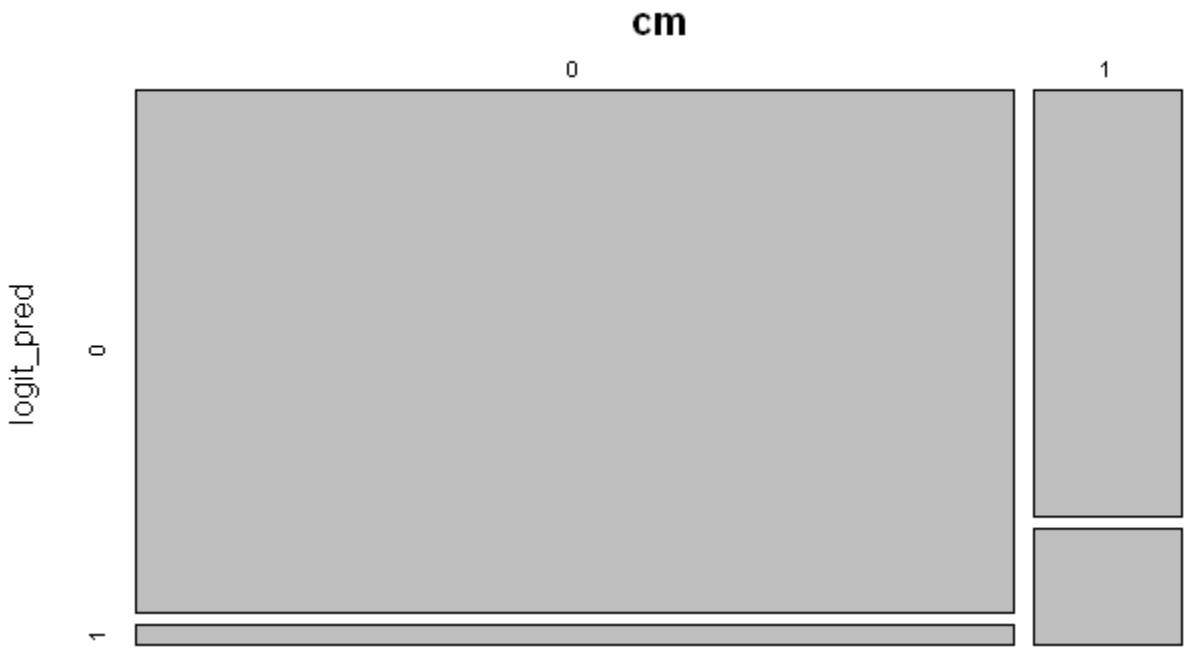
Precision is defined as the number of true positives divided by the sum of true positives and false positives, where true positive means a correctly predicted positive result, and false positive means a result predicted positive but actually negative. Precision can be thought of as “of the things we predicted positive, what fraction of those are correct?”

A recall can be thought of as “of the actual number of positive results, what fraction did we predict correctly?”. We use a small function to do these calculations. Results are shown below -

```
Accuracy -> ((TN+TP)/(TN+TP+FN+FP))*100))
precision ->  ',((TP)/(TP+FP))*100)
recall//TPR ->  ',((TP)/(TP+FP))*100)
```

```
Number of cases in table: 999
Number of factors: 2
Test for independence of all factors:
        Chisq = 64.16, df = 1, p-value = 1.145e-15

Confusion Matrix and Statistics

   logit_pred
      0   1
  0 822  33
  1 113  31

               Accuracy : 0.8539
                 95% CI : (0.8304, 0.8752)
    No Information Rate : 0.9359
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2298

 Mcnemar's Test P-Value : 6.231e-11

            Sensitivity : 0.8791
            Specificity : 0.4844
         Pos Pred Value : 0.9614
         Neg Pred Value : 0.2153
             Prevalence : 0.9359
         Detection Rate : 0.8228
   Detection Prevalence : 0.8559
      Balanced Accuracy : 0.6818

       'Positive' Class : 0

[1] "Logistic Regression Accuracy 0.817909168808912"
RESULTS
[1] "Accuracy :- 85.3853853853854"
[1] "FNR :- 78.4722222222222"
[1] "FPR :- 3.85964912280702"
[1] "precision :-  48.4375"
[1] "recall//TPR :-  48.4375"
[1] "Sensitivity :-  21.5277777777778"
[1] "Specificity :-  96.140350877193"
```

Confusion Matrix Plot for model 1

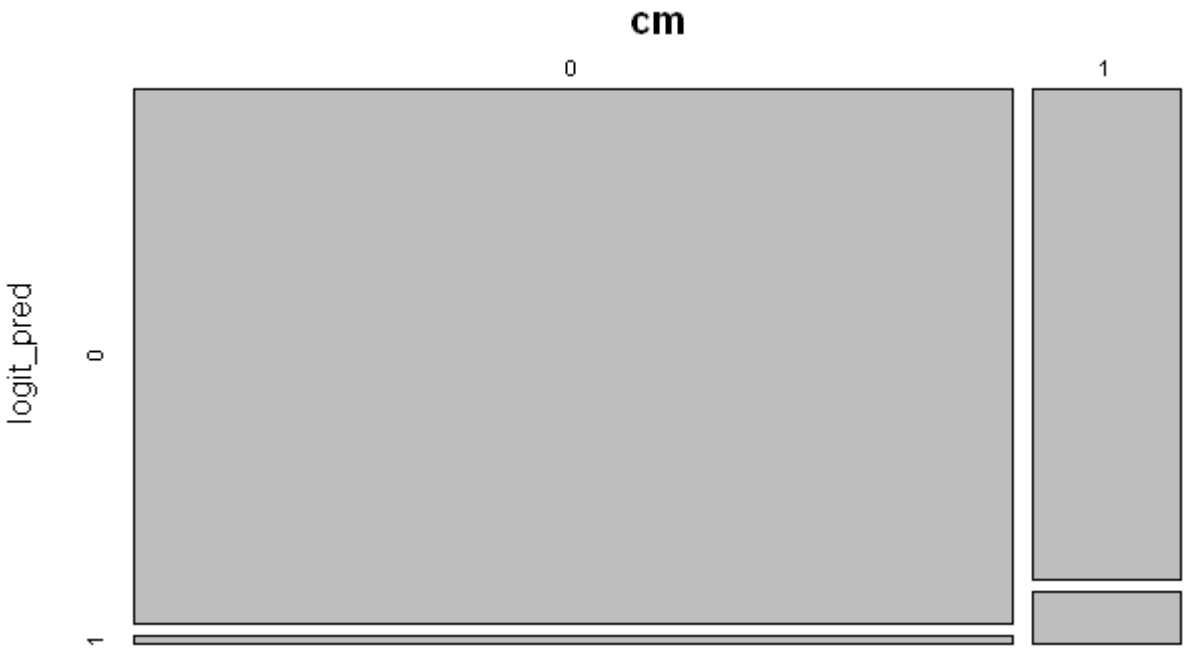


- Evaluate the performance of model 2 using Confusion Matrix. The accuracy in this model is slightly improved to 83%. Precision is improved to 50%.

```
Number of cases in table: 999
Number of factors: 2
Test for independence of all factors:
       Chisq = 29.57, df = 1, p-value = 5.393e-08
       Chi-squared approximation may be incorrect

Confusion Matrix and Statistics

   logit_pred
      0   1
  0 841  14
  1 130  14

               Accuracy : 0.8559
                 95% CI : (0.8325, 0.8771)
    No Information Rate : 0.972
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1216

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

            Sensitivity : 0.86612
```

```
            Specificity : 0.50000
         Pos Pred Value : 0.98363
         Neg Pred Value : 0.09722
             Prevalence : 0.97197
         Detection Rate : 0.84184
   Detection Prevalence : 0.85586
      Balanced Accuracy : 0.68306

       'Positive' Class : 0

[1] "Logistic Regression Accuracy 0.838474721508141"

[1] "Accuracy :- 85.5855855855856"
[1] "FNR :- 90.2777777777778"
[1] "FPR :- 1.6374269005848"
[1] "precision :-  50"
[1] "recall//TPR :-  50"
[1] "Sensitivity :-  9.72222222222222"
[1] "Specificity :-  98.3625730994152"
```

Confusion Matrix Plot for model 2



- VIF for both the models

  The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

```
> vif(logit_model)
                    GVIF Df GVIF^(1/(2*Df))
ContractRenewal    1.072580  1        1.035654
DataPlan          14.954223  1        3.867069
DataUsage       1730.690318  1       41.601566
CustServCalls      1.229486  9        1.011544
DayMins          929.234483  1       30.483348
DayCalls           1.017585  1        1.008754
MonthlyCharge   2968.181565  1       54.481020
OverageFee       214.161155  1       14.634246
RoamMins           1.221395  1        1.105167
wkcategory         1.100710  9        1.005345
```

```
> vif(logit_model2)
ContractRenewal         DayMins   MonthlyCharge      OverageFee        RoamMins
       1.019325        1.651561        1.817301        1.173858        1.018174
```

- ANOVA on both the models

ANOVA test on Predictors suggest we can leave out Overage fees, Data usage and Account Weeks from the list and proceed to build a model without these predictor variables for Logit regression.

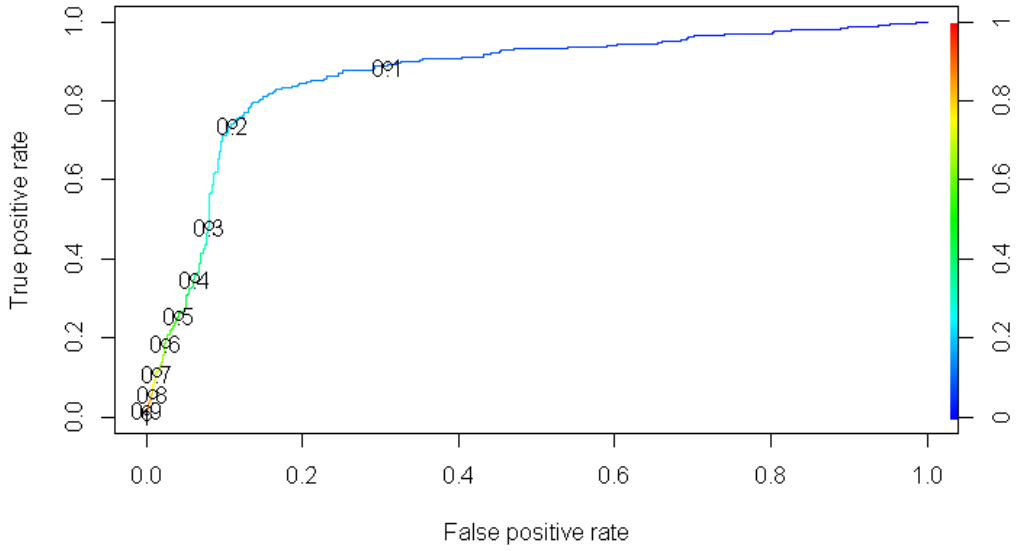
```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          2333     1934.3
ContractRenewal  1  129.883      2332     1804.4 < 2.2e-16 ***
DataPlan         1   28.236      2331     1776.2 1.074e-07 ***
DataUsage        1    4.306      2330     1771.9   0.03798 *
CustServCalls    9  199.380      2321     1572.5 < 2.2e-16 ***
DayMins          1  101.921      2320     1470.5 < 2.2e-16 ***
DayCalls         1    0.062      2319     1470.5   0.80354
MonthlyCharge    1   22.762      2318     1447.7 1.834e-06 ***
OverageFee       1    2.511      2317     1445.2   0.11304
RoamMins         1    3.229      2316     1442.0   0.07233 .
wkcategory       9    4.802      2307     1437.2   0.85125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          2333     1934.3
ContractRenewal  1  129.883      2332     1804.4 < 2.2e-16 ***
DayMins          1   82.582      2331     1721.8 < 2.2e-16 ***
MonthlyCharge    1    8.744      2330     1713.1  0.003106 **
OverageFee       1   29.652      2329     1683.4 5.171e-08 ***
RoamMins         1   10.121      2328     1673.3  0.001466 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

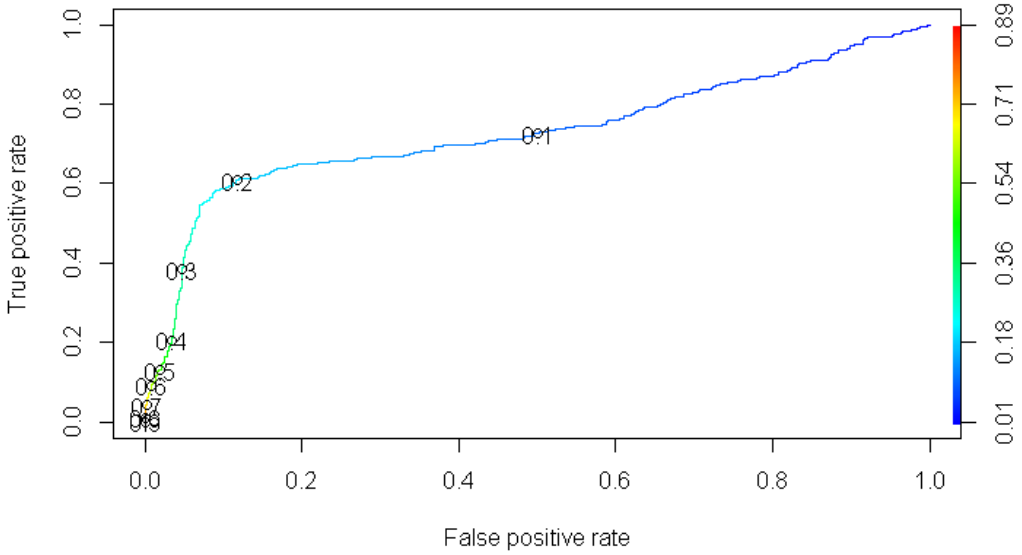
- ROC Curve

ROC curve is a plot of sensitivity (the ability of the model to predict an event correctly) versus 1-specificity for the possible cut-off classification probability values. The Area Under the Curve (AUC), also referred to as index of accuracy (A) and it is an accepted traditional performance metric for a ROC curve. The higher the area under the curve the better prediction power the model has.

Model 1 ROC curve



Model 2 ROC curve

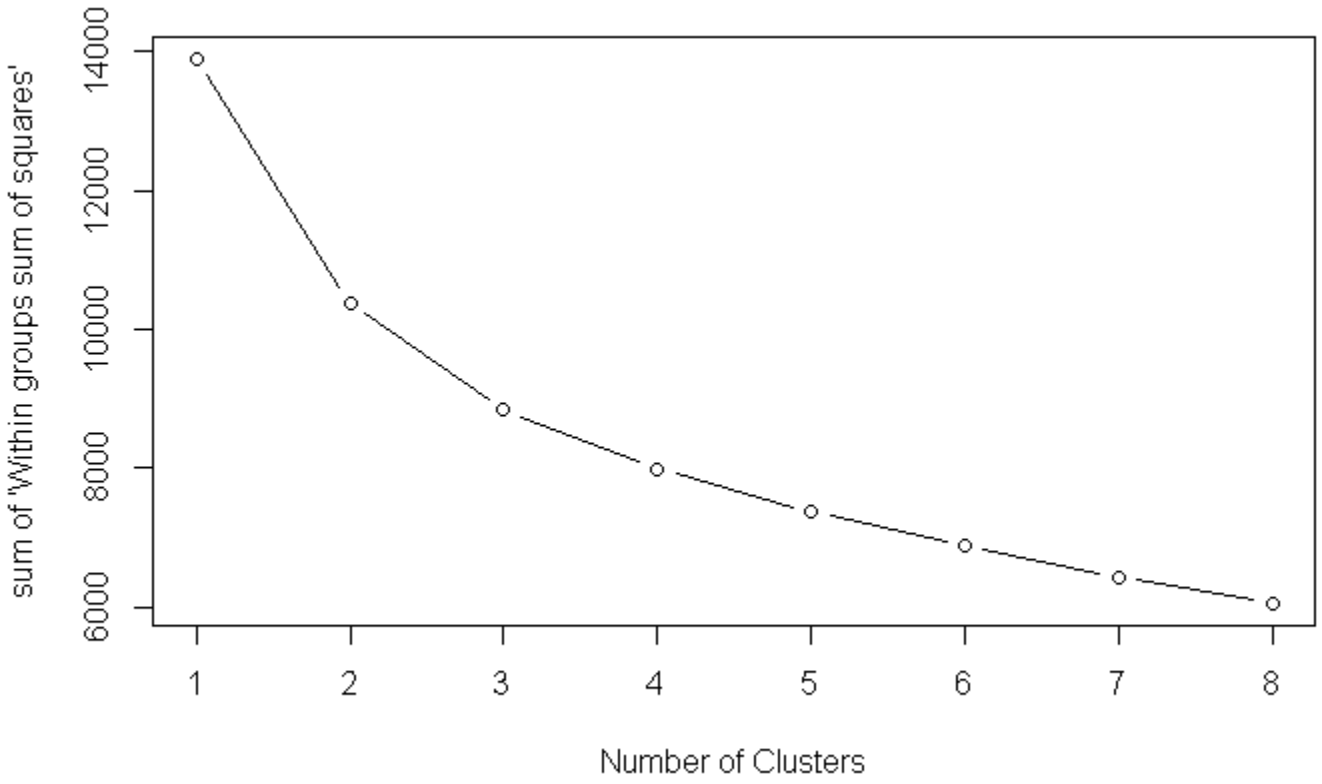


KNN CLASSIFICATION

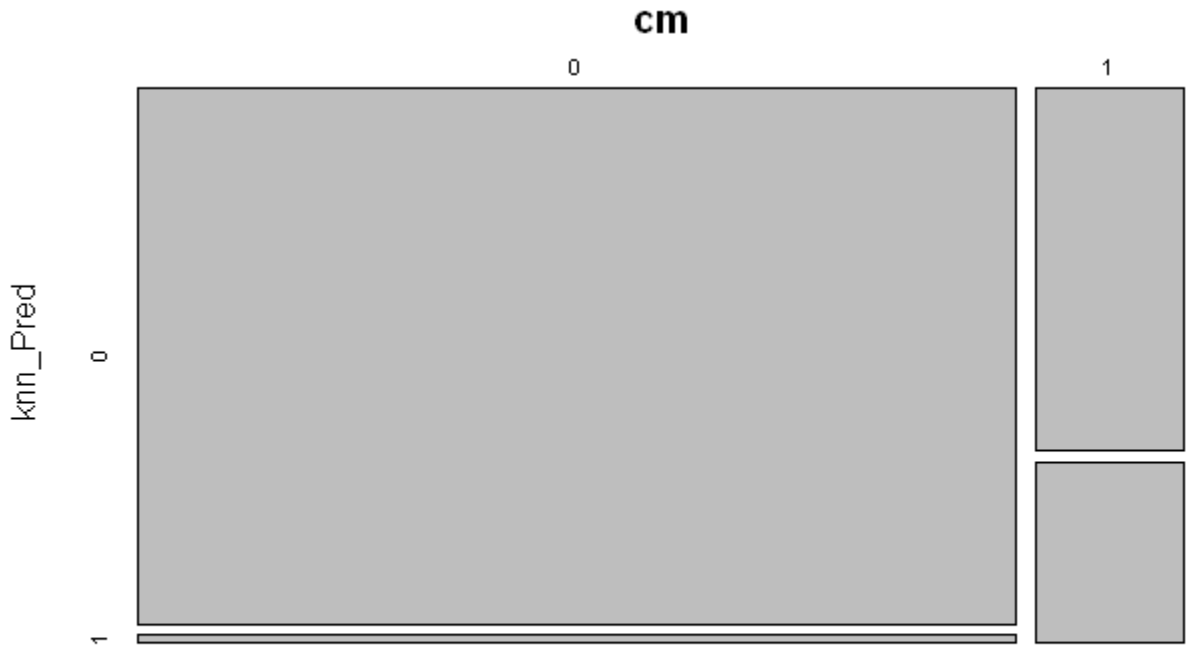
KNN classification iterations shows the best vale of k=5 gives optimum results. Accordingly, the confusion matrix is built to calculate the accuracy, precision, and recall values. Accuracy is 89% , precision and recall values are at 78% much higher than the logistic regression model.

```
k-Nearest Neighbors

2334 samples
   8 predictor
   2 classes: '0', '1'

Pre-processing: centered (8), scaled (8)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 2101, 2101, 2100, 2100, 2101, 2101, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.8928914  0.4741785
   7  0.8986010  0.4891127
   9  0.8994527  0.4813027
  11  0.8990223  0.4711364
  13  0.8983076  0.4619671
  15  0.8968788  0.4493853
  17  0.8951670  0.4416705
  19  0.8950208  0.4380163
  21  0.8935908  0.4243722
  23  0.8944473  0.4300265

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 9.

> mean(knn.pred == test.cell$Churn)
[1] 0.8878879

Confusion Matrix and Statistics

   knn_Pred
      0   1
  0 842  13
  1  96  48

               Accuracy : 0.8909
                 95% CI : (0.8699, 0.9096)
    No Information Rate : 0.9389
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4184

 Mcnemar's Test P-Value : 4.024e-15

            Sensitivity : 0.8977
            Specificity : 0.7869
```

```
         Pos Pred Value : 0.9848
         Neg Pred Value : 0.3333
             Prevalence : 0.9389
         Detection Rate : 0.8428
   Detection Prevalence : 0.8559
      Balanced Accuracy : 0.8423

       'Positive' Class : 0

> calc(cm_knn)
[1] "Accuracy :- 89.0890890890891"
[1] "FNR :- 66.6666666666667"
[1] "FPR :- 1.52046783625731"
[1] "precision :-  78.6885245901639"
[1] "recall//TPR :-  78.6885245901639"
[1] "Sensitivity :-  33.3333333333333"
[1] "Specificity :-  98.4795321637427"
```



Confusion Matrix plot for KNN model



NAIVE BAYES

Naive Bayes Classifier for Discrete Predictors

```
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace, type = "class")

A-priori probabilities:
Y
        0         1
0.8547558 0.1452442

Conditional probabilities:
   ContractRenewal
Y            0         1
  0 0.06516291 0.93483709
  1 0.29498525 0.70501475

   DataPlan
Y           0         1
  0 0.7057644 0.2942356
  1 0.8289086 0.1710914
```

```
   DataUsage
Y          [,1]      [,2]
  0  0.03012757 1.0042245
  1 -0.18926375 0.9503223

   CustServCalls
Y             0           1           2           3           4           5
     6
  0 0.2150375940 0.3754385965 0.2255639098 0.1358395990 0.0350877193 0.0075187970 0.0035
087719
  1 0.1858407080 0.2713864307 0.1828908555 0.0737463127 0.1474926254 0.0855457227 0.0353
982301
   CustServCalls
Y             7           8           9
  0 0.0015037594 0.0005012531 0.0000000000
  1 0.0147492625 0.0000000000 0.0029498525

   DayMins
Y          [,1]      [,2]
  0 -0.06814608 0.9158758
  1  0.47981632 1.2307816

   DayCalls
Y           [,1]      [,2]
  0  0.001395957 0.9711795
  1 -0.024646895 1.0865579

   MonthlyCharge
Y          [,1]      [,2]
  0 -0.02612746 1.0005633
  1  0.17722258 0.9951173

   OverageFee
Y          [,1]      [,2]
  0 -0.04269218 0.9915792
  1  0.20656609 1.0258761

   RoamMins
Y          [,1]      [,2]
  0 -0.02824217 0.9969034
  1  0.15221081 1.0133443

   wkcategory
Y             6         6-12        12-18        18-24        24-30        30-36
 36-42
  0 0.0295739348 0.0947368421 0.1669172932 0.2426065163 0.2401002506 0.1378446115 0.0641
604010
  1 0.0324483776 0.0707964602 0.1681415929 0.2359882006 0.2713864307 0.1386430678 0.0678
466077
   wkcategory
Y          42-48        48-54        54-60
  0 0.0205513784 0.0030075188 0.0005012531
  1 0.0088495575 0.0058997050 0.0000000000
```

```
> mean(NB.pred==test.cell$Churn)
[1] 0.8068068


C Confusion Matrix and Statistics

   NB.pred
      0   1
  0 833  22
  1 103  41

               Accuracy : 0.8749
                 95% CI : (0.8527, 0.8948)
    No Information Rate : 0.9369
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3381

 Mcnemar's Test P-Value : 8.342e-13

            Sensitivity : 0.8900
            Specificity : 0.6508
         Pos Pred Value : 0.9743
         Neg Pred Value : 0.2847
             Prevalence : 0.9369
         Detection Rate : 0.8338
   Detection Prevalence : 0.8559
      Balanced Accuracy : 0.7704

       'Positive' Class : 0

[1] "Accuracy :- 87.4874874874875"
[1] "FNR :- 71.5277777777778"
[1] "FPR :- 2.57309941520468"
[1] "precision :-  65.0793650793651"
[1] "recall//TPR :-  65.0793650793651"
[1] "Sensitivity :-  28.4722222222222"
[1] "Specificity :-  97.4269005847953"
```

MODEL COMPARISON USING - CONFUSION MATRIX INTERPRETATION FOR ALL MODELS

CONFUSION MATRIX	LOGISTIC REGRESSION	KNN	NAÏVE BAYES
Accuracy	85.6%	89%	87.5%
FNR	90.3%	66.7%	71.5%
PRECISION	50%	78.7%	65%
RECALL (TPR)	50%	78.7%	65%
SENSITIVITY (TNR)	9.7%	33.3%	89%
SPECIFICITY	98.3%	98.5%	65%
F-SCORE	92.1%	94%	93%

Accuracy seems to be good in all models. Based on the precision, recall value, KNN model is better than other models is classifying the cluster of customers and to evaluate the action to be taken to reduce the churn. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results.

ACTIONABLE INSIGHTS AND RECOMMENDATIONS

To guide the recommendation, let's go back and try to answer our queries about my customer segments:

- Does the number of calls made to Customer Service indicate the individuals who are more likely to churn?

  Customer Service calls appeared to be not so significant for lower call rate, but it is advised to carefully look at customers with the higher call rate to customer care.

- Do individuals with Data Plan and Data Usage more like to churn more than those without a Data Plan?
  In this particular dataset, the data plan and data usage are not significant.

- Does the number of Calls made per month or the Roaming Calls made per month significant in classifying individuals pattern?

- Does the Monthly Charges or the Overage Fees indicate the individuals who are more likely to churn?

Both the above parameters are highly significant in identifying the customer churn. It is advised to offer the customer the next best thing by making the real-time recommendations that have the greatest likelihood of acceptance to the offer inline with their usage trends.