Telecom Customer churn modeling

Table of contents

1 Project Objective

The primary objective of this report is to help the Telecom industry for the prediction of postpaid customer churn. The data has information about the customer usage behavior, contract details and the payment details. The data also indicates which were the customers who canceled their service.

2 Exploratory data analysis

Generally, Exploratory data analysis are carried out to discover the patterns in the given dataset,missing values ,central tendency of the given dataset.

2.1 Environment set up and data Import

The working directory should be one , where the code and dataset are placed.

The following R packages used for the analysis of Churn prediction dataset

- Ggplot2 package-To infer about the attributes of the loan prediction data using graphical plots
- caTools-To split the given dataset into training and testing
- corrplot-To plot the correlations between predictors
- Class-To build the K- Nearest Neighbours model
- e1071-Naïve bayes model building
- ROCR- To compute model performance measures such as KS,Area under curve, lift chart
- Ineq-To compute the gini index of data
- InformationValue-Package to calculate the concordance ratio

2.2 Variable Identification:

Some basics R functions such as mean , sd ,round are used for the statistical calculation. Here are the following functions are getting used for the better understanding of data and to make further decision **.**

Structure(Str)- To get the structure of the customer churn prediction such as class category ,name and the count of the fields

Summary –Generally , summary function will give the classification of attributes. Here, the customer churn prediction data having Eleven columns/ fields such as Account weeks, Contract Renewal, Data Plan, Data usage ,Monthly charge etc .If the field is categorical data, summary function give the count of each sub-category If the field is integer, then it will return the five stats notably minimum,25th percentile , median or 50th percentile , 75th percentile and maximum.

2.3 Uni-Variate Analysis:

Summary statistics

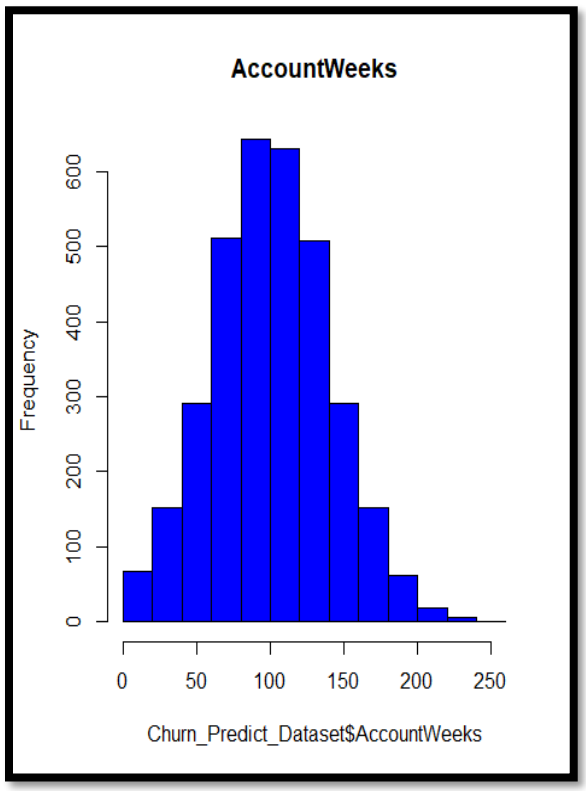
```
> summary(Churn_Predict_Dataset)
 Churn      AccountWeeks   ContractRenewal DataPlan   DataUsage      CustServCalls      DayMins          DayCalls
 0:2850   Min.   :  1.0   0: 323          0:2411   Min.   :0.0000   Min.   :0.000   Min.   :  0.0   Min.   :  0.0
 1: 483   1st Qu.: 74.0   1:3010          1: 922   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:143.7   1st Qu.: 87.0
          Median :101.0                            Median :0.0000   Median :1.000   Median :179.4   Median :101.0
          Mean   :101.1                            Mean   :0.8165   Mean   :1.563   Mean   :179.8   Mean   :100.4
          3rd Qu.:127.0                            3rd Qu.:1.7800   3rd Qu.:2.000   3rd Qu.:216.4   3rd Qu.:114.0
          Max.   :243.0                            Max.   :5.4000   Max.   :9.000   Max.   :350.8   Max.   :165.0
 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
```

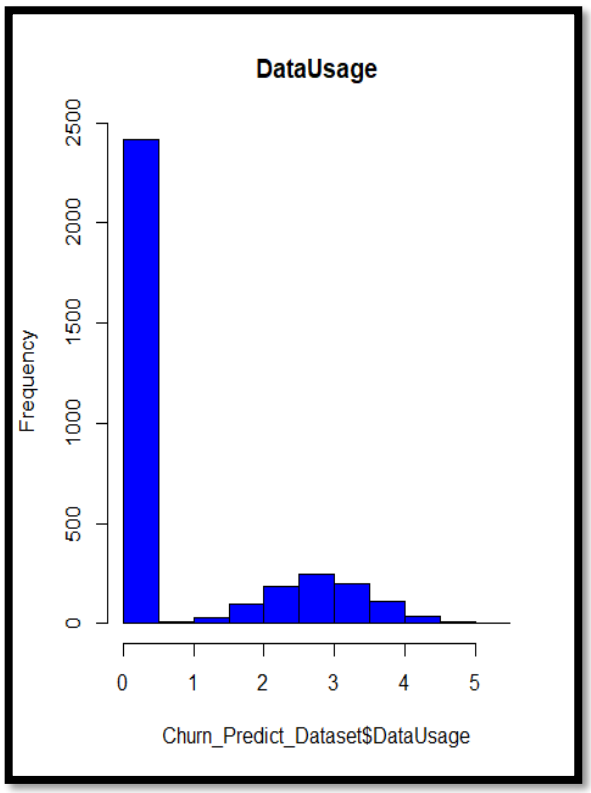
Frequency distribution of dependent variable

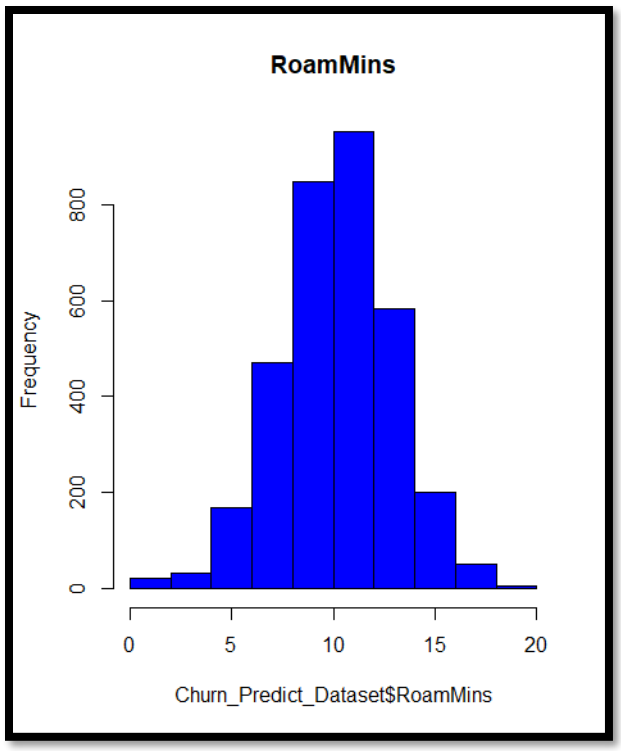
Category	Count
Customer Churn(0)	2850
Customer Churn(1)	483
Total	3333

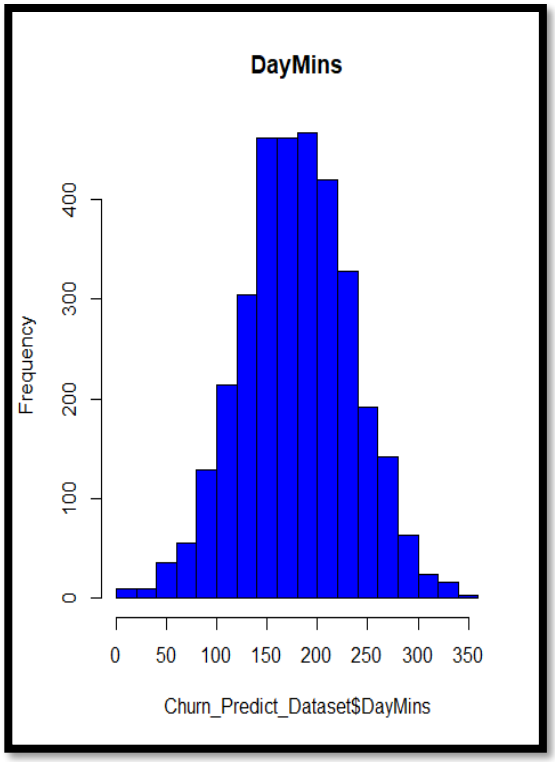
Among these 3333 customers, 14 percentage of customers are already cancelled their postpaid service.

Frequency distribution of independent variables



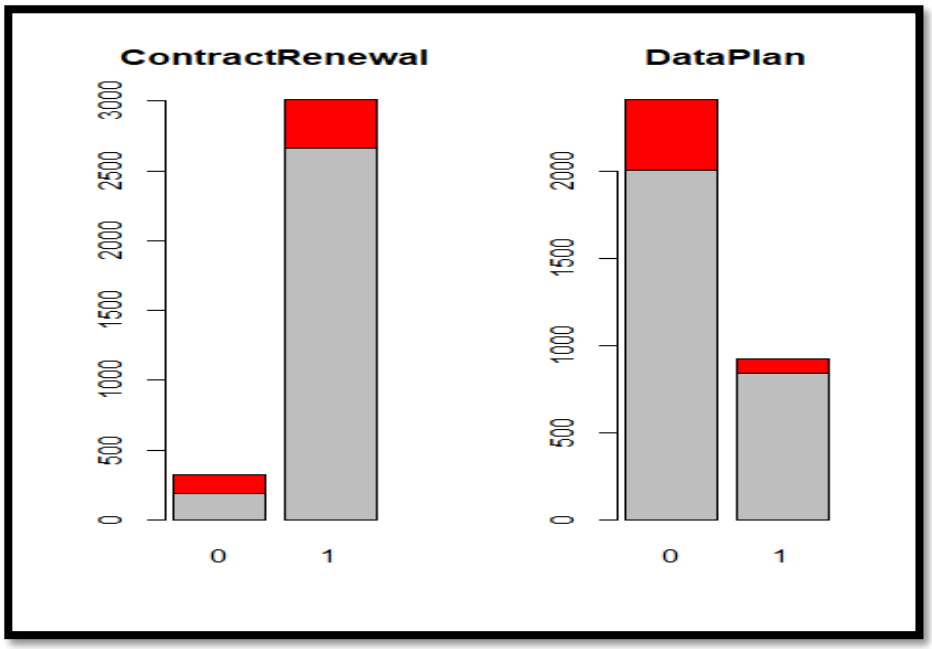






2.4 Bi-Variate Analysis:

Relationship between Categorical IV's and target variable



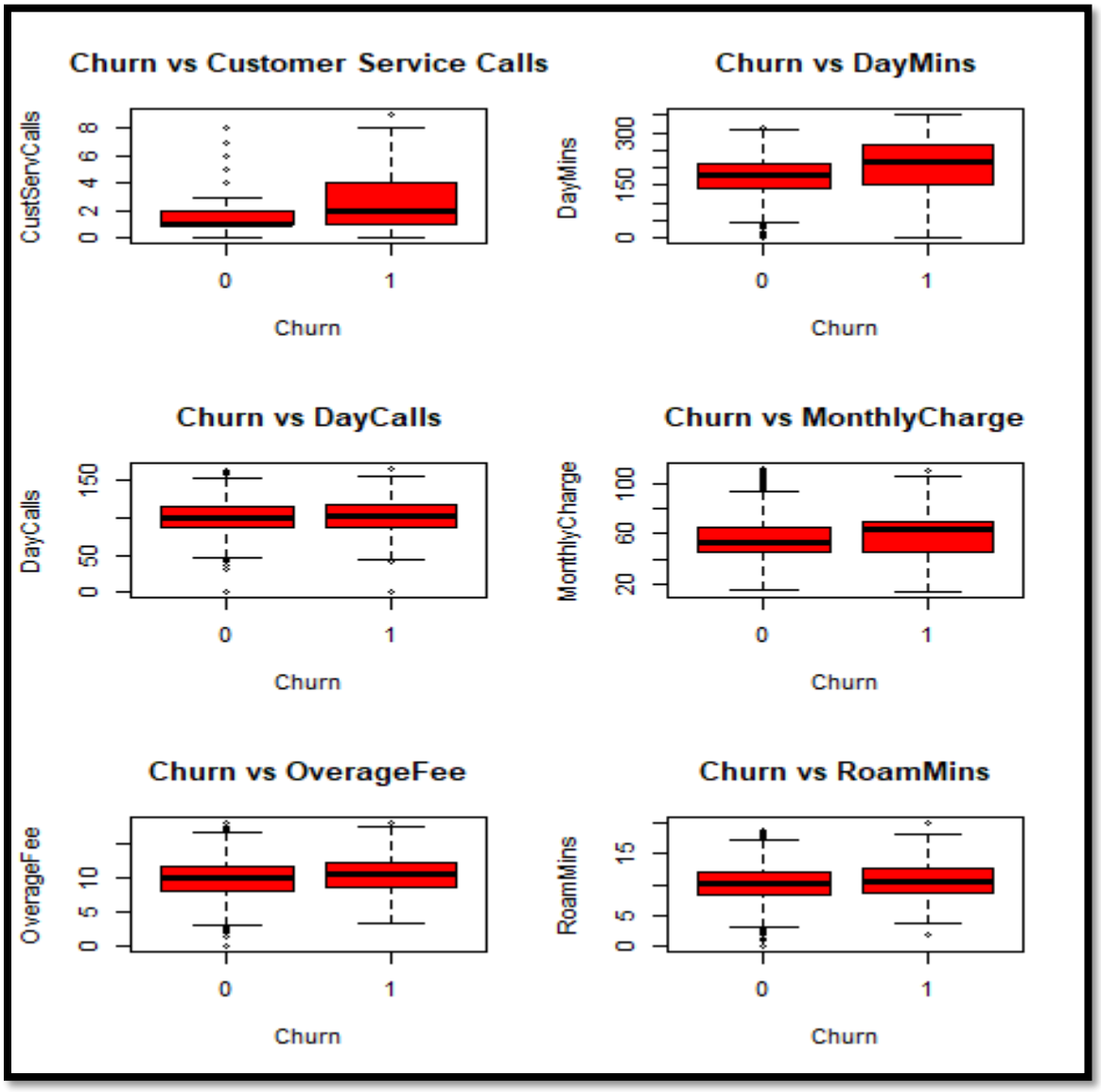
The above plots will give the relationship between the categorical IV(Independent Variable) and target variable(ie. Churn vs contract renewal and data plan).

Red color - Customer who cancelled services(ie churn=1)

Gray color - Customer who still using postpaid service(ie churn=0)

Most of the customers have done the contract renewal, so the data is more biased towards 1(people who did the contract renewal).Due to bias, this variable is statistically not significant.

Relationship between Predictors and Target variable

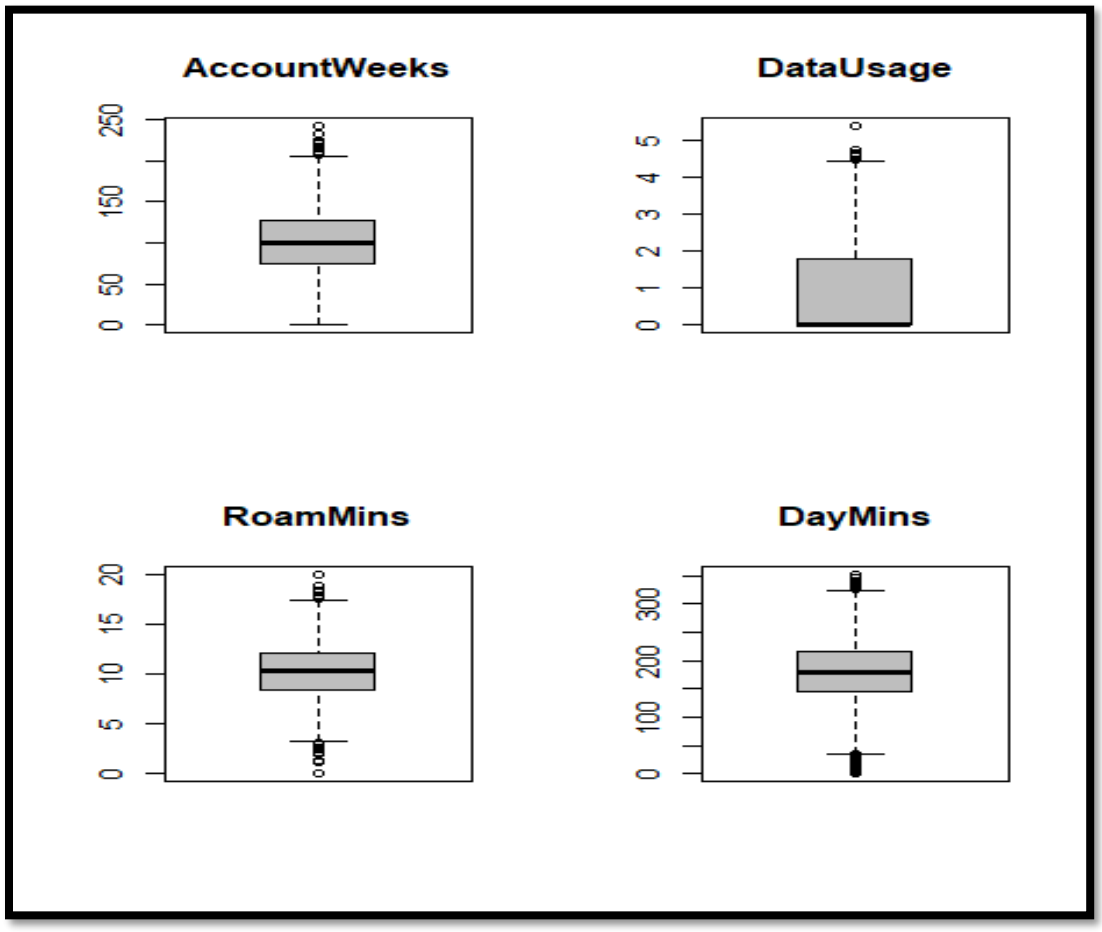


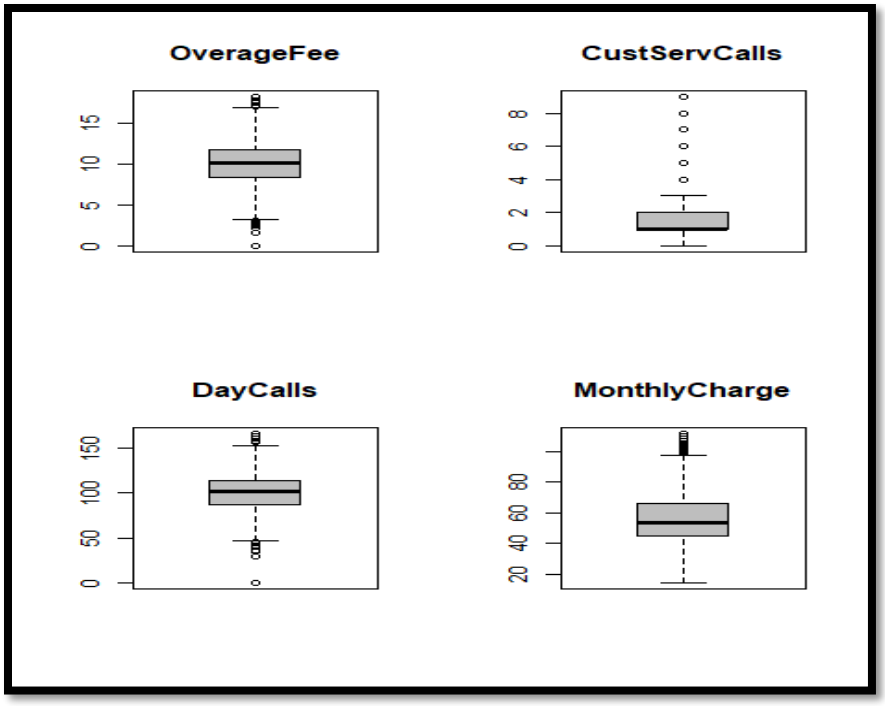
Missing values and Outliers :

Currently 3333 customers are using telecom postpaid service .There is no missing values in the given dataset.

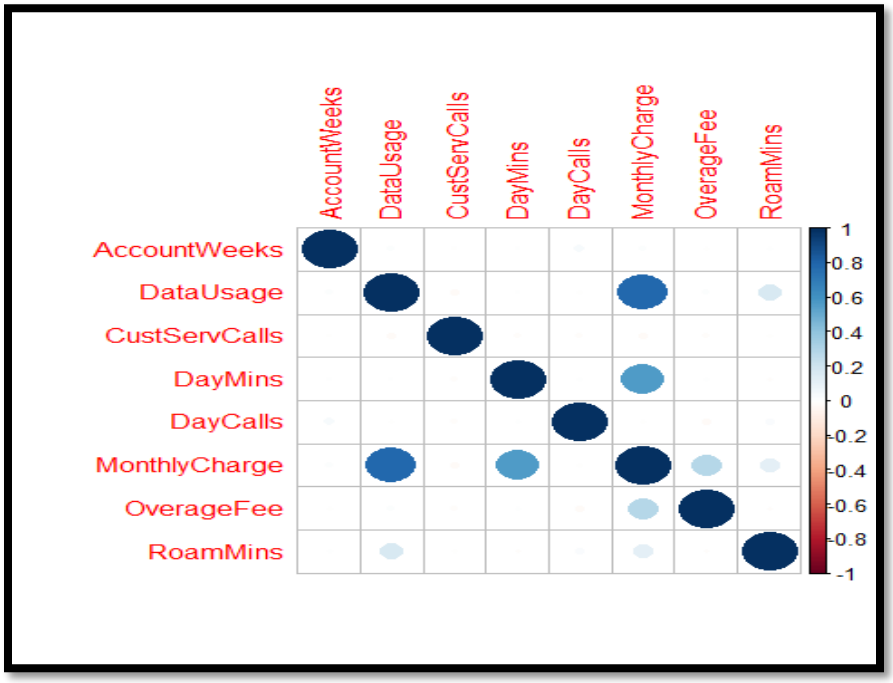
```
> #Checking for missing values
> sum(is.na(Churn_Predict_Dataset))#no missing value
[1] 0
> ##  check that no datapoint is missing, otherwise we need to fix the dataset.
> apply(Churn_Predict_Dataset,2,function(x) sum(is.na(x)))
         Churn    AccountWeeks ContractRenewal        DataPlan       DataUsage   CustServCalls         DayMins
             0               0               0               0               0               0               0
     DayCalls   MonthlyCharge      OverageFee        RoamMins
             0               0               0               0
>
```

From the boxplot ,it is infer that there is an outlier in the given dataset.





Multi-Collinearity:



The Independent variables which are highly correlated are refered to as multi-collinearity problem. Following methods can be used to detect multicollinearity:

- The analysis exhibits sign of multicollinearity - such as, coefficient estimates varying from model to model
- The R-square value is large but none of the beta weights is statistically significant, i.e., F-test for overall model is significant but the t-tests for individual coefficient estimates are not.
- Correlation among pairs of variables are large.
- Variance Inflation Factor.

In the given customer churn dataset ,data usage and monthly charge are highly correlated.Based on the p-value ,monthly charge field is less significant and removed from building the model .

Inferences from EDA:

To Check the visual association pattern for continous predictor variables with target variable.

One would expect a decreasing churn rate with the increase in the time (account weeks) of an account, but it does not seem to be the case.Clearly, there is a good probability (approx 40%) of an account churning if the contract has not been renewed

The churn rate increases if a customer makes 4 or more calls to the customer service and also churn rate increases if the monthly average daytime minutes are greater than 245.

3)Supervised Learning Techniques:

3.1)Logistic Regression Model:

Logistic regression can be used for both classification and regression models.Mostly, it will be used for classification problems and because of that it is known as Binomial classifier.It is the base model for the classification technique.

glm is generally used to fit generalized linear models.However, in this case, you need to make it clear that you want to fit a logistic regression model, resolve this by setting the family argument to binomial.

```
##################Logistics regression#################
#Individual relationship between Predictors and  Target variable
summary(glm(Churn_Predict_Train$Churn~Churn_Predict_Train$AccountWeeks,data =Churn_Predict_Train,family ="binomial" ))#Insignificant
summary(glm(Churn_Predict_Train$Churn~Churn_Predict_Train$DayCalls,data =Churn_Predict_Train,family ="binomial" ))#Insignificant

summary(glm(Churn_Predict_Train$Churn~Churn_Predict_Train$ContractRenewal,data =Churn_Predict_Train,family ="binomial" ))#significant
summary(glm(Churn_Predict_Train$Churn~Churn_Predict_Train$DataPlan,data =Churn_Predict_Train,family ="binomial" ))#significant
summary(glm(Churn_Predict_Train$Churn~Churn_Predict_Train$DataUsage,data =Churn_Predict_Train,family ="binomial" ))#significant
summary(glm(Churn_Predict_Train$Churn~Churn_Predict_Train$CustServCalls,data =Churn_Predict_Train,family ="binomial" ))#significant
summary(glm(Churn_Predict_Train$Churn~Churn_Predict_Train$DayMins,data =Churn_Predict_Train,family ="binomial" ))#significant
summary(glm(Churn_Predict_Train$Churn~Churn_Predict_Train$MonthlyCharge,data =Churn_Predict_Train,family ="binomial" ))#significant
summary(glm(Churn_Predict_Train$Churn~Churn_Predict_Train$OverageFee,data =Churn_Predict_Train,family ="binomial" ))#significant
summary(glm(Churn_Predict_Train$Churn~Churn_Predict_Train$RoamMins,data =Churn_Predict_Train,family ="binomial" ))#significant
```

p value determines the probability of significance of predictor variables. With 95% confidence level, a variable having $p < 0.05$ is considered an important predictor.The predictors such as account weeks and daycalls are insignificant when compared to individual relationship between other predictor variables.

Building the model with all predictors:

Summary() function returns the he estimate, standard errors, z-score, and p-values on each of the coefficients.Look like most of the co-efficients are not significant here while building the model with all predictors.

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0370  -0.5135  -0.3403  -0.2016   3.0398

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -6.0247220  0.6454540  -9.334  < 2e-16 ***
AccountWeeks      0.0013174  0.0016805   0.784    0.433
ContractRenewal1 -2.0138901  0.1696443 -11.871  < 2e-16 ***
DataPlan1        -1.3998621  0.6527393  -2.145    0.032 *
DataUsage         1.4951928  2.2876705   0.654    0.513
CustServCalls     0.5278694  0.0472797  11.165  < 2e-16 ***
DayMins           0.0349165  0.0386171   0.904    0.366
DayCalls          0.0009193  0.0032518   0.283    0.777
MonthlyCharge    -0.1369380  0.2267900  -0.604    0.546
OverageFee        0.3901690  0.3867789   1.009    0.313
RoamMins          0.1132337  0.0260700   4.343  1.4e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1930.4  on 2332  degrees of freedom
Residual deviance: 1517.9  on 2322  degrees of freedom
AIC: 1539.9
```

Variation inflation factor(VIF):

```
> vif(Log_Model_2)
   AccountWeeks ContractRenewal        DataPlan       DataUsage   CustServCalls         DayMins        DayCalls   MonthlyCharge
       1.004884        1.071782       14.216113     1584.430645        1.086480      964.354648        1.002382     2850.679951
     OverageFee       RoamMins
     207.970352        1.185902
>
```

Eliminating the insignificant variables with the help of vif values.Data usage ,monthly charge and average fee are having high vif value.

Build the model for Significant variables:

Regression equation is log odds(y) =(-5.9013 -2.0281 * Contract Renewal-1.0390*DataPlan +0.5267*CustServCalls+0.1567*OverageFee+0.1194*RoamMins+0.0116*DayMins)

```
Call:
glm(formula = Churn_Predict_Train$Churn ~ ContractRenewal + DataPlan +
    CustServCalls + OverageFee + RoamMins + DayMins, family = "binomial",
    data = Churn_Predict_Train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0465  -0.5121  -0.3416  -0.2014   3.0357

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.901314   0.515903 -11.439  < 2e-16 ***
ContractRenewal1 -2.028122   0.169184 -11.988  < 2e-16 ***
DataPlan1       -1.039021   0.175069  -5.935 2.94e-09 ***
CustServCalls    0.526765   0.047100  11.184  < 2e-16 ***
OverageFee       0.156762   0.027290   5.744 9.23e-09 ***
RoamMins         0.119461   0.024198   4.937 7.94e-07 ***
DayMins          0.011639   0.001265   9.201  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1930.4  on 2332  degrees of freedom
Residual deviance: 1519.3  on 2326  degrees of freedom
AIC: 1533.3

Number of Fisher Scoring iterations: 6

> vif(Log_Model_3)
ContractRenewal        DataPlan    CustServCalls      OverageFee        RoamMins         DayMins
       1.066694        1.026795        1.084260        1.036562        1.018345        1.036279
```

The logistic regression coefficients give the change in the log odds of the outcome for a unit increase in the predictor variable. For every unit increase/decrease in predictors , the log odds of being Churn = 1 increase/decrease by predictor cofficients.

Odds Ratio:

Exponentiate the coefficients and interpret them as odds-ratio.

```
> print(exp(Log_Model_3$coefficients))
     (Intercept) ContractRenewal1        DataPlan1    CustServCalls       OverageFee         RoamMins          DayMins
     0.002735848      0.131582450      0.353801003      1.693445780      1.169717250      1.126888732      1.011707095
>
```

Prediction in Logistic Regression:

Given dataset was getting splitted in the ratio of 70 and 30.Predicting the results for the 30% of given dataset.Response’s are decided based on probability threshold value.Initially the threshold value was set as 0.50.Even it shows good of amount accuracy,the given data is imbalanced data.So we can’t conclude the performance of model only with accuracy.

Sensitivity of the logistic model is 0.17

```
> Churn_Predict_Test$Prob_test=predict(Log_Model_3,newdata=Churn_Predict_Test,type ="response")
> Churn_Predict_Test$Predicted_data=ifelse(Churn_Predict_Test$Prob_test<0.50,0,1)
> Accuracy_lg=mean(Churn_Predict_Test$Predicted_data==Churn_Predict_Test$Churn)#imbalanced data,we can't conclude w
h accuracy
> Accuracy_lg
[1] 0.859
>
```

Find the optimal threshold value to increase the sensitivity by decreasing the threshold. By decreasing the probability threshold value, there is increase in the True positive rate / Sensitivity of the model. TPR increased from 0.17 to 0.52.

3.2)K-Nearest Neighbours Model:

KNN is a Supervised Learning algorithm that uses labeled input data set to predict the output of the data points. It is mainly based on feature similarity. KNN checks how similar a data point is to its neighbor and classifies the data point into the class it is most similar to.

Knn() function-The mandatory paramters are

1)Training dataset

2)Testing dataset

3)Target variable

4)K-Value

Optimal value of k:

K values are calculated from the input features. One of the ways to find the optimal K value is to calculate the square root of the predictors in the data set.

Total number of input variables=11

Square root of 11 is 3.31.So the optimal k-value is 3

#Load class package to build KNN model

library(class)

Churn_Predict_Test_KNN$Status=knn(Churn_Predict_Train_KNN[,-c(1)],Churn_Predict_Test_KNN[,-c(1)],Churn_Predict_Train_KNN$Churn,k=3)

Euclidean distance is used to measure the similarities between the predictors.For the classification problem,majority vote determine the response classes.

```
> #Confusion matrix
> Confusion_matrix_KNN=table(Churn_Predict_Test_KNN$Churn,Churn_Predict_Test_KNN$Status)
> Confusion_matrix_KNN

     0   1
  0 818  37
  1 104  41
> Accuracy_KNN=(Confusion_matrix_KNN[1,1]+Confusion_matrix_KNN[2,2])/1000#imbalanced data,
racy
> Accuracy_KNN
[1] 0.859
>
```

From the confusion matrix , it is inferred that KNN model predicted 41 customers correctly from 3333 customers where customer likely to became a churn.

3.3)Naïve Baye's Model:

Naïve Baye's algorithm is basically derived from Baye's theorem,Where the classification of responses are based on conditional probability.Prior probability and evidence will be given,condition is to find the posterior probability.

Model Assumptions:

Naïve Bayes assumes all the features to be conditionally independent. So, if some of the features are in fact dependent on each other (in case of a large feature space), the prediction might be poor.

In other words, it assumes that the presence of one feature in a class is completely unrelated to the presence of all other features. If this assumption of independence holds, Naive Bayes performs extremely well and often better than other models.

Naïve Baye's for continuous variables:

Naive Bayes can also be used with continuous features but is more suited to categorical variables. If all the input features are categorical, Naive Bayes is recommended. However, in case of numeric features, it makes another strong assumption which is that the numerical variable is normally distributed.

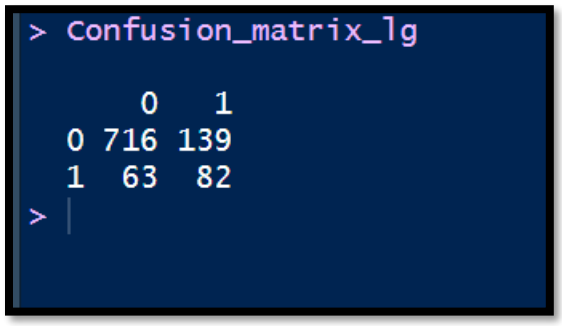
Interpretation of Naïve Baye's:

The features of the customer churn dataset is highly dependent to each other. So we can't build the naïve bayes model.

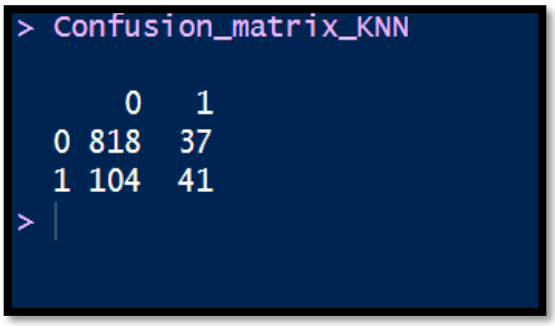
4)Model Performance Metrics:

Confusion matrix

Logistic regression



KNN Model



Logistics regression:

The test data consisted of 1000 observations. Out of which 82 cases have been accurately predicted (TN->True Negatives) as Churn in nature .Also, 716 out of 1000 observations were accurately predicted (TP-> True Positives) as not churn in nature which constitutes 71%.

There were no cases of False Negatives (FN) meaning no cases were recorded which actually are churn in nature but got predicted as not churn.

There were 63 cases of False Positives (FP) meaning 63 cases were actually not churn in nature but got predicted as churn.

The total accuracy of the model is 79.8 %((TN+TP)/1000) which shows that there may be chances to improve the model performance

KNN Model:

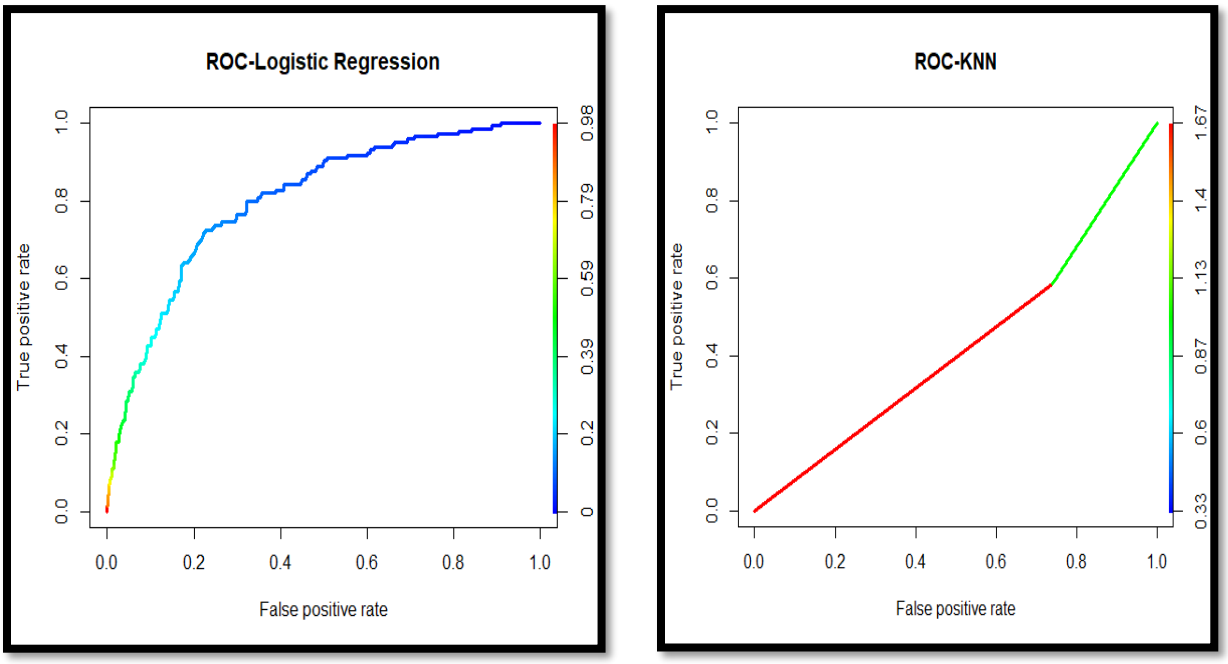
The test data consisted of 1000 observations. Out of which 41 cases have been accurately predicted (TN->True Negatives) as Churn in nature .Also, 818 out of 1000 observations were accurately predicted (TP-> True Positives) as not churn in nature which constitutes 81%.

There were no cases of False Negatives (FN) meaning no cases were recorded which actually are churn in nature but got predicted as not churn.

There were 104 cases of False Positives (FP) meaning 104 cases were actually not churn in nature but got predicted as churn.

The total accuracy of the model is 85.9 %((TN+TP)/1000) which shows that there may be chances to improve the model performance.

ROC and Area under Curve:



The ROC curve is the plot between sensitivity and (1- specificity). (1- specificity) is also known as false positive rate and sensitivity is also known as True Positive rate.

Roc of Logistic regression is giving enough information about the model when compared to KNN.

KS,AUC and Gini Index:

	TPR	FPR	Accuracy	KS	AUC	Gini
Logistic Regression Model	56.5%	83.7%	79.8%	0.49	82%	52.9
KNN Model	28%	95.6%	85.9%	0.1	42%	0.16

Based on the model performance ,Logistic regression is best model when compared to the KNN.Eventhough accuracy is low on logistic model,but it has high sensitivity rate .

Deciles and Rank ordering:

```
#3Deciling code-Rank ordering
qs_lg=quantile(Churn_Predict_Test$Prob_test,prob = seq(0,1,length=11))
print(qs_lg)
print(qs_lg[10])
threshold=qs_lg[10]
mean((Churn_Predict_Test$Churn[Churn_Predict_Test$Prob_test>threshold])=="1")
Churn_Predict_Test$Deciles=cut(Churn_Predict_Test$Prob_test,unique(qs_lg),include.lowest = TRUE,right = FALSE)
head(Churn_Predict_Test)

#Rank ordering
library(data.table)
#Loan_Dataset_lg$Personal.Loan=as.numeric(Loan_Dataset_lg$Personal.Loan)
DT_lg=data.table(Churn_Predict_Test)
#Aggregate columns

Rtable_lg=DT_lg[,list(cnt=length(Churn),
                      cnt_tar1 = sum(Churn==1),
                      cnt_tar0 = sum(Churn==0)),by=Deciles][order(-Deciles)]
print(Rtable_lg)
Rtable_lg$rrate = round(Rtable_lg$cnt_tar1 / Rtable_lg$cnt,4)*100;
Rtable_lg$cum_resp = cumsum(Rtable_lg$cnt_tar1)
Rtable_lg$cum_non_resp = cumsum(Rtable_lg$cnt_tar0)
Rtable_lg$cum_rel_resp = round(Rtable_lg$cum_resp / sum(Rtable_lg$cnt_tar1),4)*100;
Rtable_lg$cum_rel_non_resp = round(Rtable_lg$cum_non_resp / sum(Rtable_lg$cnt_tar0),4)*100;
Rtable_lg$ks = abs(Rtable_lg$cum_rel_resp - Rtable_lg$cum_rel_non_resp);
print(Rtable_lg)

# Concordance Function
library(InformationValue)
Concordance(actuals=Churn_Predict_Test$Churn, predictedScores=Churn_Predict_Test$Prob_test)
```

Conclusion:

Naïve Baye's cant be used for the given dataset,due to its dependencies between predictors.KNN model will give better responses when all the variables are categorical ,but the given customer churn is collection of both categorical and continuous predictors.So , the KNN is not performed well as compared to the rest of the models.Logistic regression is used for binomial classification ,so this model predicts the customer who likely to became churn in the future.