Telecom customer churn Prediction assessment

K Vineet PATNAIK

Table of contents

EDA

There are different libraries used in the following project which are mentioned in the code .

CODE

```
library(readxl)
library(plyr)
getwd()
setwd("C:/Users/vineet patnaik/Desktop/R language/text files/")
mydata = read_excel("Cellphone.xlsx")
mydata                      # shows my data
dim(mydata)                 # shows total observations(3333) and variables (11)
names(mydata)               #gives all the names of the 11 variables
attach(mydata)
count(mydata[1:3333,],vars = "Churn")  Churn  freq
                                       1      0 2850
                                       2      1  483
2850/(483+2850)  ## 85.55% where churn=0 or customer is staying
```

```
str(mydata)  #structure of the data
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 ...
```

Univariate Analysis

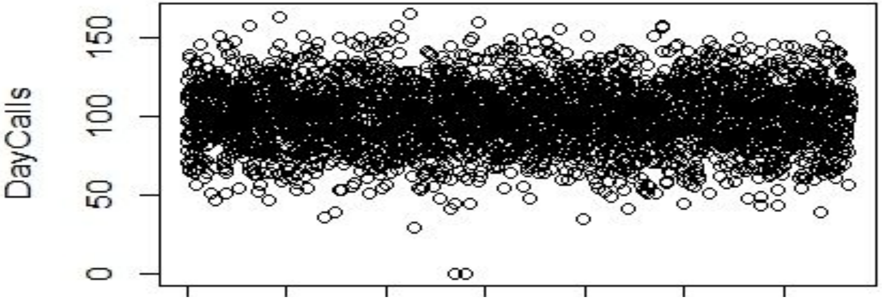
Univariate is basically the descriptive analysis of a single variable .this can be done by the summary of the data

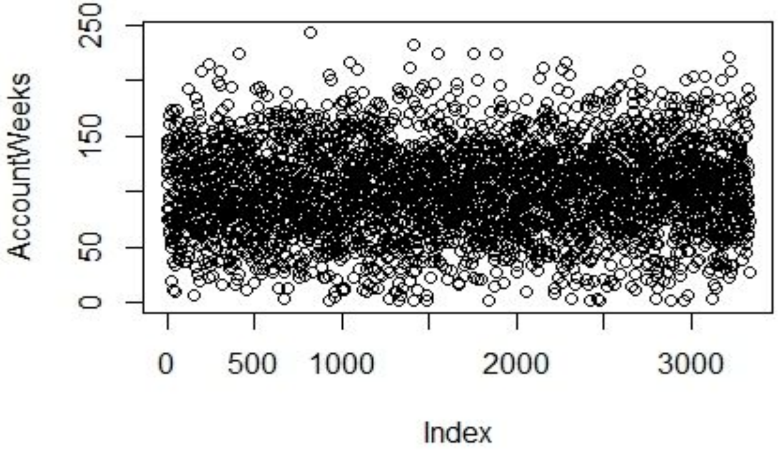
summary(mydata)

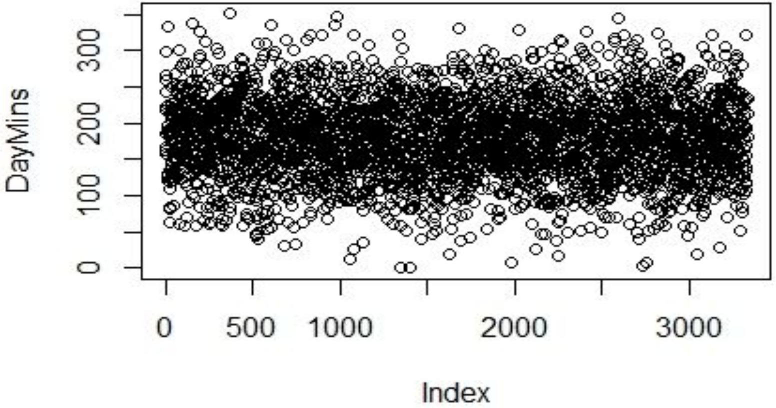
```
> summary(mydata)
     Churn         AccountWeeks    ContractRenewal     DataPlan         DataUsage       CustServCalls       DayMins     
 Min.   :0.0000   Min.   :  1.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :  0.0  
 1st Qu.:0.0000   1st Qu.: 74.0   1st Qu.:1.0000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

Churn is a discrete variable with churn 0 = 2850,churn 1 = 483 and the remaining are continuous variables with min ,median, mean and max values given along with different quartile values

Plots & graphs







Bivariate analysis

Correlation.. ..Statistical test to determine the strength of a relationship between two variables. It is between -1 to +1 ,where close to either positive or negative 1 is strong relationship and a value close to zero is a weaker relationship .

```
library(caret)
correlationMatrix <- cor(mydata[,1:11])
print(correlationMatrix)
highlyCorrelated <- findCorrelation(correlationMatrix, cutoff=0.5)
print(highlyCorrelated)
```

	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins
Churn	1.00000000	0.016540742	-0.259851847	-0.102148141	-0.087194509	0.208749999	0.205150829
AccountWeeks	0.01654074	1.000000000	-0.024734655	0.002918409	0.014390757	-0.003795939	0.006216021
ContractRenewal	-0.25985185	-0.024734655	1.000000000	-0.006006371	-0.019222913	0.024521956	-0.049395824
DataPlan	-0.10214814	0.002918409	-0.006006371	1.000000000	0.945981734	-0.017823944	-0.001684069
DataUsage	-0.08719451	0.014390757	-0.019222913	0.945981734	1.000000000	-0.021722518	0.003175951
CustServCalls	0.20875000	-0.003795939	0.024521956	-0.017823944	-0.021722518	1.000000000	-0.013423186
DayMins	0.20515083	0.006216021	-0.049395824	-0.001684069	0.003175951	-0.013423186	1.000000000
DayCalls	0.01845931	0.038469882	-0.003754626	-0.011085902	-0.007962079	-0.018941930	0.006750414
MonthlyCharge	0.07231271	0.012580670	-0.047291399	0.737489653	0.781660429	-0.028016853	0.567967924
OverageFee	0.09281243	-0.006749462	-0.019104644	0.021525559	0.019637372	-0.012964219	0.007038214
RoamMins	0.06823878	0.009513902	-0.045870743	-0.001317871	0.162745576	-0.009639680	-0.010154586

	DayCalls	MonthlyCharge	OverageFee	RoamMins
Churn	0.018459312	0.072312711	0.092812426	0.068238776
AccountWeeks	0.038469882	0.012580670	-0.006749462	0.009513902
ContractRenewal	-0.003754626	-0.047291399	-0.019104644	-0.045870743
DataPlan	-0.011085902	0.737489653	0.021525559	-0.001317871
DataUsage	-0.007962079	0.781660429	0.019637372	0.162745576
CustServCalls	-0.018941930	-0.028016853	-0.012964219	-0.009639680
DayMins	0.006750414	0.567967924	0.007038214	-0.010154586
DayCalls	1.000000000	-0.007963218	-0.021448602	0.021564794
MonthlyCharge	-0.007963218	1.000000000	0.281766048	0.117432607
OverageFee	-0.021448602	0.281766048	1.000000000	-0.011023336
RoamMins	0.021564794	0.117432607	-0.011023336	1.000000000

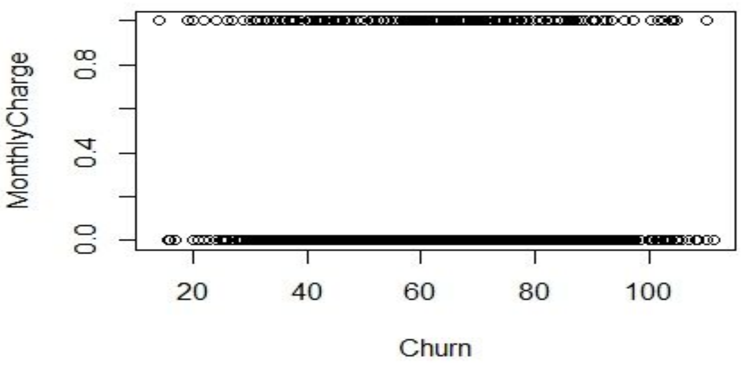
Here the highest correlation values are 0.94(dataplan,datausage) ,0.73(monthlycharge,dataplan),0.78(monthlycharge,datausage),0.56(monthly charge,datamins)

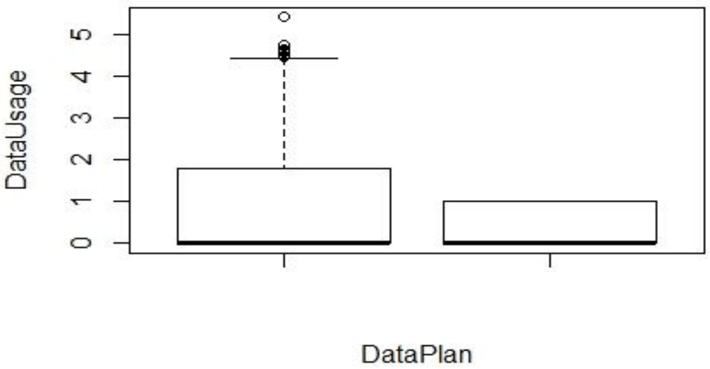
Linear regression

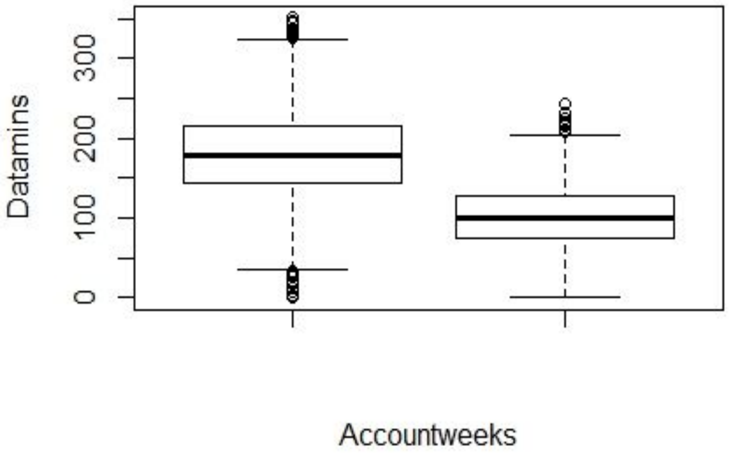
```
summary(lm(Churn~AccountWeeks))
summary(lm(Churn~DayMins))
summary(lm(Churn~ContractRenewal))
summary(lm(Churn~DayCalls))
summary(lm(Churn~DataPlan))
summary(lm(Churn~MonthlyCharge))
summary(lm(Churn~OverageFee))
summary(lm(Churn~CustServCalls))
summary(lm(Churn~RoamMins))
summary(lm(Churn~DataUsage))

## they explain how much there linear relationship between the variables is and run just lm too
```

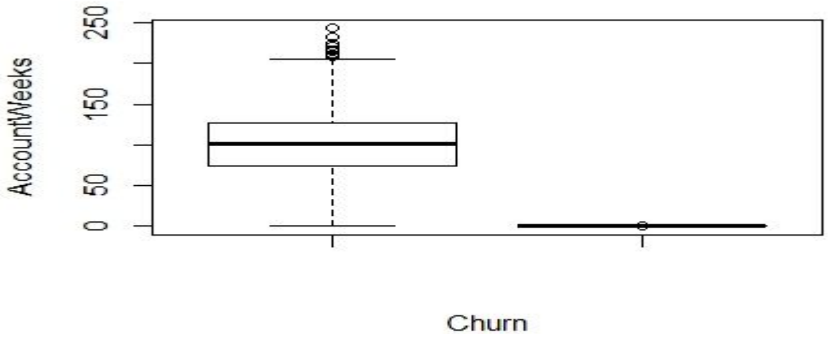
Plots & graphs

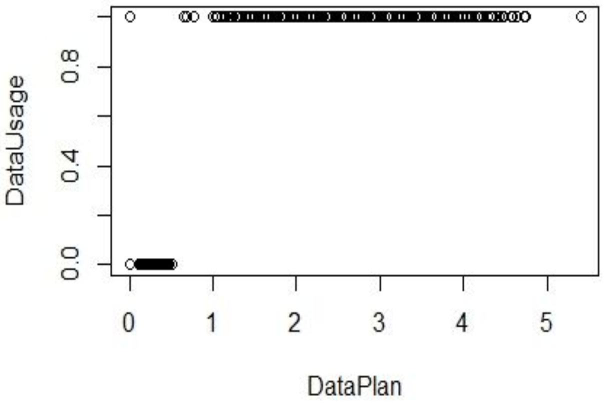
```
boxplot(AccountWeeks,Churn,xlab = "Churn",ylab = "AccountWeeks")
boxplot(DataUsage,DataPlan,ylab = "DataUsage",xlab = "DataPlan")
boxplot(DayMins,AccountWeeks,ylab = "Datamins",xlab ="AccountW)
boxplot(DayMins,Churn,xlab = "Churn",ylab = "DayMins")
plot(MonthlyCharge,Churn,xlab = "Churn",ylab = "MonthlyCharge")
plot(DataUsage,DataPlan,ylab = "DataUsage",xlab = "DataPlan")
```







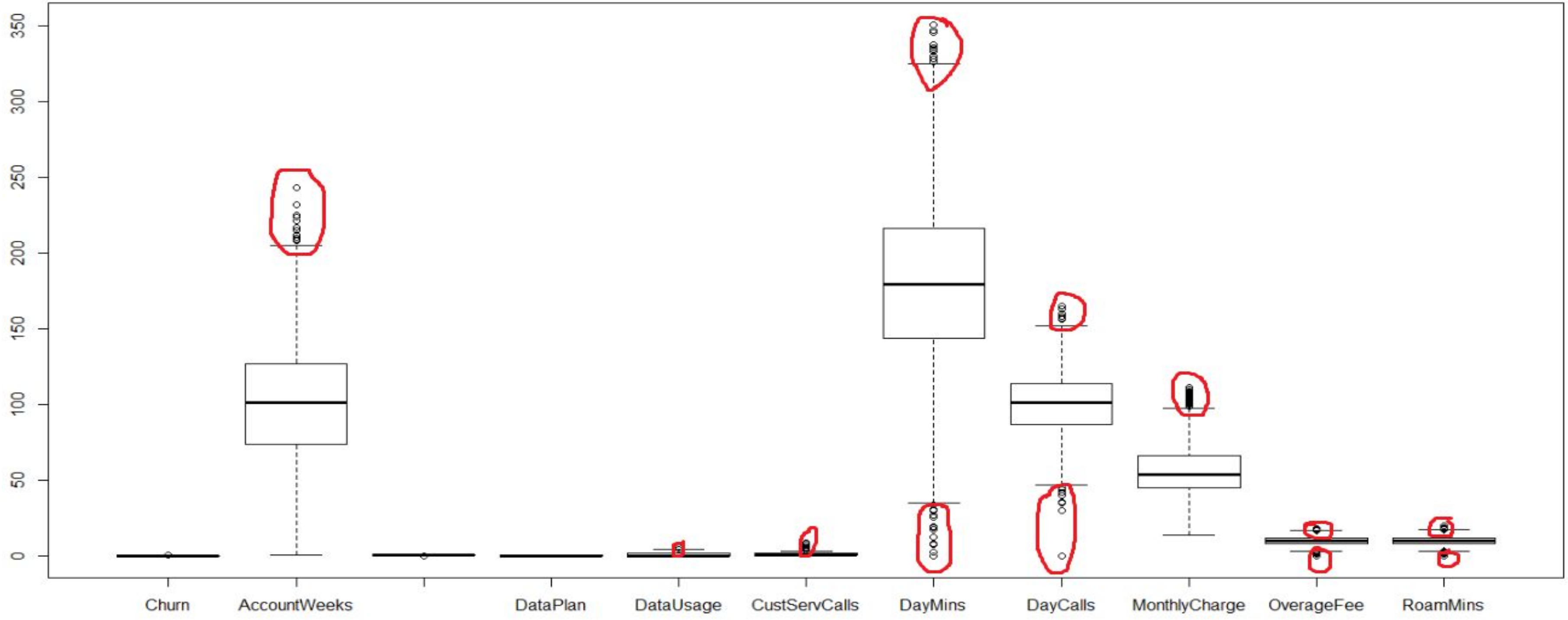




These are the boxplots and graphs between different variables

Outliers & missing values

To find the outliers in the variable we plot a boxplot and see if there are any values beyond the plot .circles ones are outliers of the data .



```
is.na(mydata)
sum(is.na(mydata))
[1] 0  ## no missing values in the data
any(is.na(mydata))
```

Multicollinearity

```
d = subset(mydata,select = -c(MonthlyCharge))
mul1 =
lm(Churn~AccountWeeks+ContractRenewal+DataPlan+DataUsage+CustSer
vCalls+DayMins+OverageFee+RoamMins,data = mydata)
summary(mul1)
vif(mul1)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -1.433e-01  5.363e-02  -2.672 0.007580 ** 
AccountWeeks     8.888e-05  1.396e-04   0.637 0.524402    
ContractRenewal -2.993e-01  1.882e-02 -15.904  < 2e-16 ***
DataPlan        -4.175e-02  4.381e-02  -0.953 0.340650    
DataUsage       -2.835e-02  1.933e-01  -0.147 0.883401    
CustServCalls    5.829e-02  4.222e-03  13.804  < 2e-16 ***
DayMins          1.021e-03  3.272e-03   0.312 0.754936    
DayCalls         3.409e-04  2.769e-04   1.231 0.218433    
MonthlyCharge    1.428e-03  1.924e-02   0.074 0.940838    
OverageFee       1.046e-02  3.280e-02   0.319 0.749780    
RoamMins         8.765e-03  2.307e-03   3.800 0.000147 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3203 on 3322 degrees of freedom
Multiple R-squared:  0.1747,	Adjusted R-squared:  0.1722 
F-statistic: 70.31 on 10 and 3322 DF,  p-value: < 2.2e-16

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

d1 = subset(mydata,select = -c(MonthlyCharge,DataUsage))
mul2 =
lm(Churn~AccountWeeks+ContractRenewal+DataPlan+CustServCalls+DayMins+OverageFee+RoamMins,data = mydata)
summary(mul2)
vif(mul2)

First we consider the multiple linear regression and take the summary of the model to understand which variables are considered , according to the data are contractrenewal,roammins,Custservcalls .then vif or variance inflation factor is taken if vif>5 then there is high multicollinearity so we have treat it by removing one of the variable with high multicollinearity which is monthlycharge ,then take a subset and run to then dataplan and

```
> vif(mul1)
  AccountWeeks ContractRenewal        DataPlan       DataUsage   CustServCalls         DayMins      OverageFee
      1.002033        1.006514       12.469512       12.813441        1.001416        1.003289        1.001214
      RoamMins
      1.345899
```

datausage are above 5 so remove datausage and check the vif

```
> vif(mul2)
   AccountWeeks ContractRenewal         DataPlan   CustServCalls         DayMins      OverageFee        RoamMins
       1.000778        1.006141         1.000815        1.001299        1.002819        1.001202        1.002555
>
```

Now there is no multicollinearity so treated completely .

Summarizing the EDA

Reading of the file is done and then the data is studied by various functions and summary is done. Then univariate and bivariate analysis is done based on the correlation and linear regression.plots and graphs are plotted for the relation with the variables.

Then outliers are identified and there are no missing values in the data then we checked for multicollinearity and treated it.

Logistic Regression

```
library(caret)
library(car)
glm(Churn~DayMins,data = mydata,family = "binomial")
# shows the linear relationship b/w two vaiables
summary(glm(Churn~DayMins,data = mydata,family = "binomial"))
```

```
Coefficients:
             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 ***
---
```

```
## reject the null hypothesis <(2e-16) therefore daymins is a significant predictor of the data
summary(glm(Churn~DayMins + AccountWeeks,data = mydata,family = "binomial"))   # check with account weeks not a good predictor
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0465610  0.2409861 -16.792   <2e-16 ***
DayMins       0.0112720  0.0009753  11.557   <2e-16 ***
AccountWeeks  0.0011517  0.0012625   0.912    0.362
---
```

summary(glm(Churn~DayMins + AccountWeeks + MonthlyCharge,data = mydata,family = "binomial"))## here the importance of accountweeks drops

summary(glm(Churn~DayMins + AccountWeeks + MonthlyCharge + CustServCalls,data = mydata,family = "binomial"))

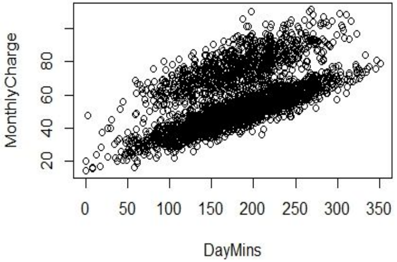
```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.684355   0.281445 -16.644   <2e-16 ***
DayMins       0.014233   0.001243  11.449   <2e-16 ***
AccountWeeks  0.001142   0.001304   0.875   0.3814
MonthlyCharge -0.012288  0.004111  -2.989   0.0028 **
CustServCalls  0.429145   0.036073  11.897   <2e-16 ***
---
```

vif(glm(Churn~DayMins + AccountWeeks + MonthlyCharge + CustServCalls ,data = mydata,family = "binomial"))

DayMins	AccountWeeks	MonthlyCharge	CustServCalls
1.505123	1.000371	1.505255	1.01876

there is no multicollinearity so we can further proceed

> cor(DayMins,MonthlyCharge)[1] 0.5679679



if we have a high correlation b/w 2 variable then we remove one and check with each variable with which the other factor importance increases if not these two variables get cancelled with each other

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.872471	0.230960	-21.10	<2e-16 ***
DayMins	0.012123	0.001011	11.99	<2e-16 ***
CustServCalls	0.430229	0.036061	11.93	<2e-16 ***

Since the null hypothesis can be rejected we can consider DayMins and CustServCalls as the two factors for logistic regression.

logisticstatus = glm(Churn~DayMins + CustServCalls,data = mydata,family = "binomial")

Logisticstatus (Intercept)	DayMins	CustServCalls
-4.87247	0.01212	0.43023

exp(0.43) = 1.53 ##which means for every month of custservcalls the odds of people not cancelling will increase by 53%

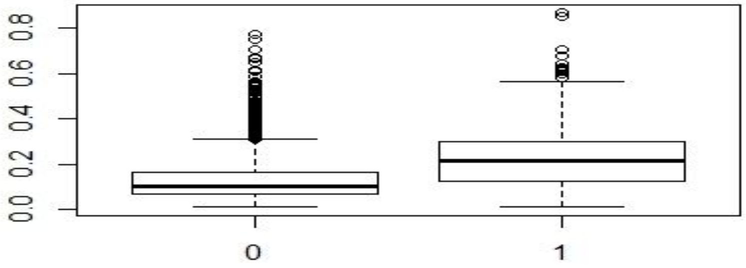
mydata$Churn = factor(mydata$Churn)

boxplot(mydata$Churn,DayMins) ## this shows more the daymins more is the customer not to cancel the service of the company

logisticstatus$fitted.values

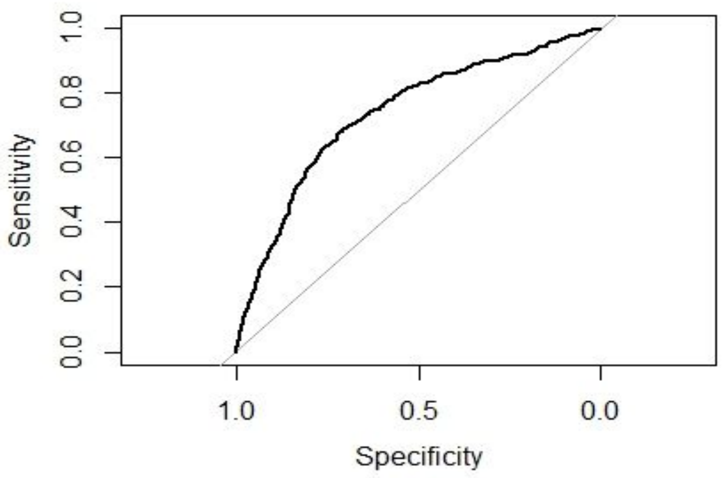
plot(mydata$Churn,logisticstatus$fitted.values) ## gives threshold values for churn=0 or 1 (probability that customer stays or quits)

[1]0.19



statuspredicted = ifelse(logisticstatus$fitted.values<0.18,"0","No")

```
table(mydata$Churn,statuspredicted)
summary(mydata$Churn)
sensitivity = 2246/2850 ## 78.7 %
sensitivity
specificity = 281/483   ## 58.3 %
specificity
(2246+281)/(3333)     ## 75..8 %
library(pROC)
roc(mydata$Churn,logisticstatus$fitted.values)
plot.roc(mydata$Churn,logisticstatus$fitted.values) ## roc curve with high sensitivity and high specificity is the best curve
```



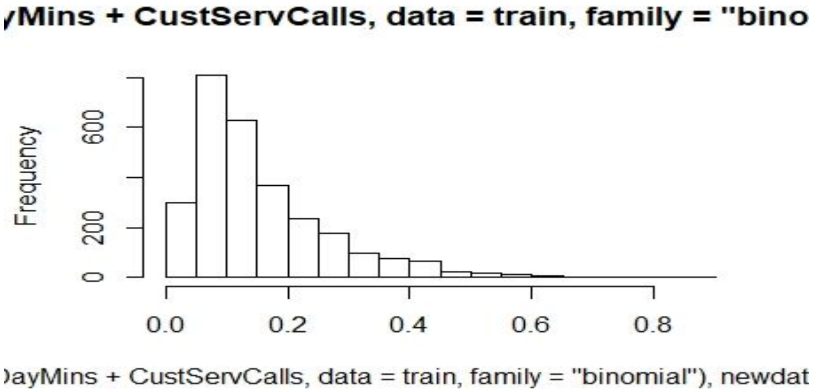

```
## there is a prediction that 75.8 % chance that indicate there will be churn=0 than 1 according to logistic regression
```

KNN

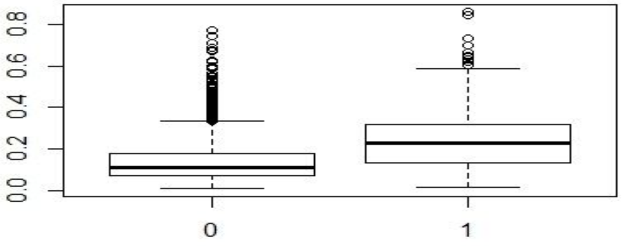
```
library(pROC)
library(ROCR)
KNN.Churn1=knn(train[,c(4,6)],test[,c(4,6)],train$Churn,k=5) ## 5 closest neighbors
KNN.Churn1
KNN.Churn=knn(scale(train[,c(4,6)]),scale(test[,c(4,6)]),train$Churn,k=21) ## needs scaling as the distance parameters tend to incline towards one variable so knn requires scaling and tuning
KNN.Churn
table(test$Churn,KNN.Churn1)
## there are 2446 samples correctly identified and 387 samples misidentified for k=5 for churn =0
## 84.8 % correct classification and 15.2 % misclassification
```

	KNN.Churn1	
	0	1
0	2351	79
1	308	95

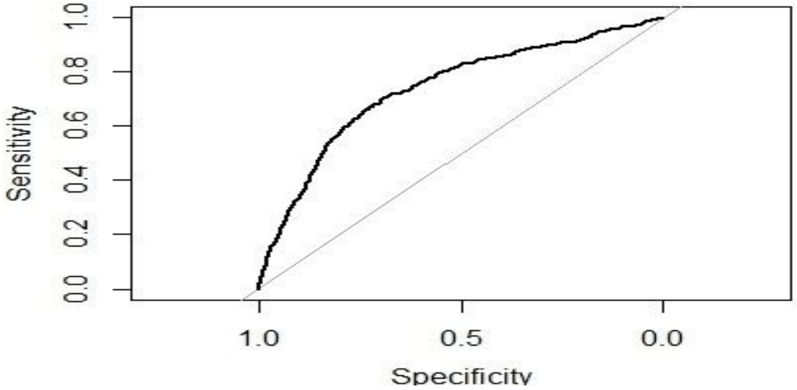
```
summary(KNN.Churn1)
   0    1
2659  174
glm(Churn~DayMins  + CustServCalls,data = train,family = "binomial")
predict(glm(Churn~DayMins+CustServCalls,data = train,family = "binomial")
,newdata = test,type = "response")
## response is used for scaling
hist(predict(glm(Churn~DayMins +  CustServCalls,data = train,family =
"binomial"),newdata = test,type = "response"))
```



it is a plot of test set predicted against a model built on train set and deployed on test set




```
roc(test$Churn,predict(glm(Churn~DayMins  + CustServCalls,data = train,family = "binomial"),newdata = test,type = "response"))
# 73.67 % area under the curve
plot(roc(test$Churn,predict(glm(Churn~DayMins + CustServCalls,data = train,family = "binomial"),newdata = test,type = "response")))
```



Naive Bayes

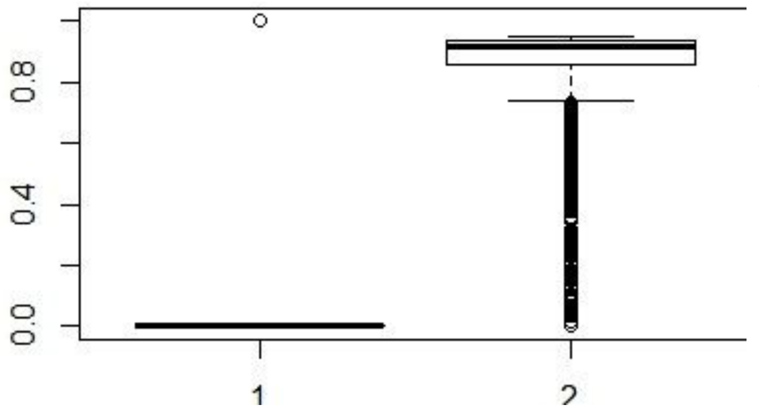
```
library(e1071)
help("naiveBayes")
naiveBayes(Churn~DayMins+CustServCalls,data = mydata)
```

```
        0         1
0.8550855 0.1449145

Conditional probabilities:
   DayMins
Y       [,1]     [,2]
  0 175.1758 50.18166
  1 206.9141 68.99779

   CustServCalls
Y       [,1]     [,2]
  0 1.449825 1.163883
  1 2.229814 1.853275
```

The conditional probabilities given by naive bayes

here 175.17 is the mean and 50.18 is the std deviance for churn =0 similarly for churn =1 for Daymins and CustServ

```
NB.churn = naiveBayes(Churn~DayMins+CustServCalls,data = mydata)
predict(NB.churn,type = "raw",newdata = mydata)
## raw is for posterior results
boxplot(Churn,predict(NB.churn,type = "raw",newdata = mydata)[,1])
predict(NB.churn,newdata = mydata)
```



Naive bayes classifier doesnt work well with this type of data since the input variables present in the data are not categorical
it works as multiclass classifier but data should be categorical since it is used for making predictions and forecasting data based on historical results
so it works well with lda or linear discriminant analysis

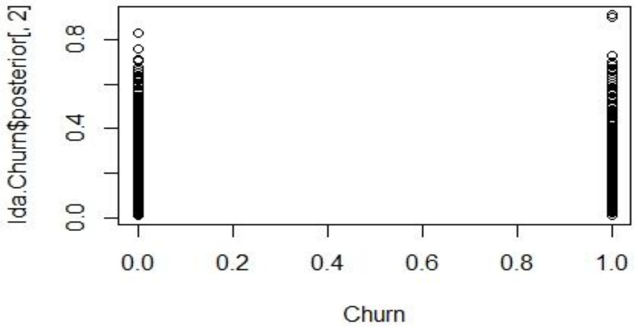
LDA

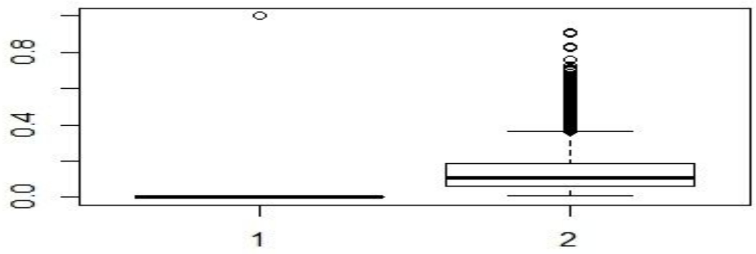
library(MASS)
library(pROC)
help(lda)
lda.Churn=lda(Churn~DayMins+CustServCalls+contractrenewal+roammins+monthlycharge, data = mydata,CV=TRUE)
lda(Churn~DayMins+CustServCalls+Contractrenewal+roammins+monthlycharge,data = mydata)

#group means show the probability that customer churn =0 or 1 w.r.t the variables

lda.Churn$posterior[,2]
plot(Churn,lda.Churn$posterior[,2])
boxplot(Churn,lda.Churn$posterior[,2])
Churn.predicted=ifelse(lda.Churn$posterior[,2]<0.95, "1","No")

```
        0         1
0.8550855 0.1449145

Group means:
   DayMins CustServCalls ContractRenewal RoamMins MonthlyCharge
0 175.1758      1.449825       0.9347368 10.15888      55.81625
1 206.9141      2.229814       0.7163561 10.70000      59.19006

Coefficients of linear discriminants:
                        LD1
DayMins          0.01207429
CustServCalls    0.44955221
ContractRenewal -2.34075379
RoamMins         0.06880425
MonthlyCharge   -0.01308329
```

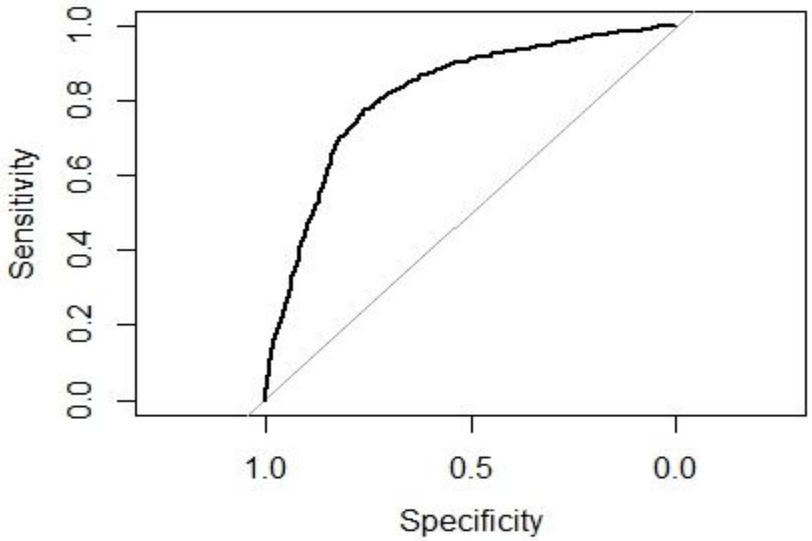




it is a bayesian method
which calculates posterior probabilities
table(Churn,Churn.predicted)

	Churn.predicted	
Churn	0	1
0	2849	1
1	477	6

plot.roc(Churn,lda.Churn$posterior[,2])
roc(Churn,lda.Churn$posterior[,2])
##Area under the curve: 0.8177



It can predict 81.77% of the time correctly which is a very good sign of classification and only 18.23% misclassification happens.
We can use class by taking 0.5 probability for both churn=0 and 1.

```
lda.Churn$class
table(Churn,lda.Churn$class)
lda(Churn~DayMins+CustServCalls+contractrenewal+roammins+monthlycharge,,prior=c(0.5,0.5),data=mydata)
lda(Churn~DayMins+CustServCalls+contractrenewal+roammins+monthlycharge,prior=c(0.5,0.5),data=mydata,CV = TRUE)
lda.Churn$posterior
boxplot(Churn,lda.Churn$posterior[,2])
table(Churn,lda.Churn$class)
```

Churn	0	1	
0	2721	129	improved class for churn=1 as previously it was 6 now 116
1	367	116	are correctly classified

Confusion matrix table

Logistic regression table

statuspredicted

	0	No
0	2246	604
1	202	281

The identifying accuracy for logistic regression is around 78.8% prediction for successful classification .

KNN

KNN.Churn1

	0	1
0	2393	30
1	366	44

The identifying accuracy for knn is around 73..3% prediction for successful classification and couldn’t predict where churn=1 very less accurate.

LDA

Churn	0	1
0	2721	129
1	367	116

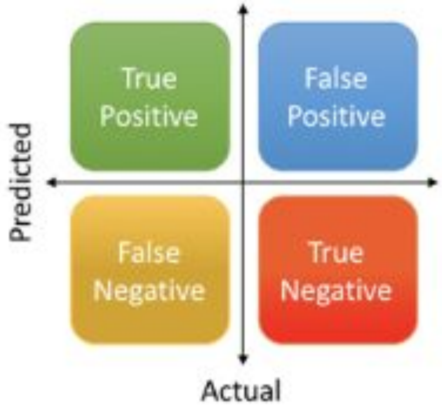
The identifying accuracy for logistic regression is around 85.11% prediction for successful classification although prediction for churn=1 is a bit less accurate.

Model Performance Metrics

Accuracy: total correct predictions/total predictions

Logistic Regression=78.8%
Knn = 73.3%
LDA = 85.11
So according to accuracy based on confusion matrix LDA is a better model.

Precision: TP/(TP+FP)
Logistic Regression = 0.782
KNN = 0.987
LDA = 0.954



So knn performs a test and training set so it has a higher Precision.but also LDA is also very good in precision

Recall:TP/(TP+FN)
Logistic Regression = 0.917
KNN = 0.867
LDA = 0.8811

So logistic regression performs well in recall since higher the recall better the model also LDA and KNN performs well too it refers to relevant results correctly classified by the model .also known as sensitivity .

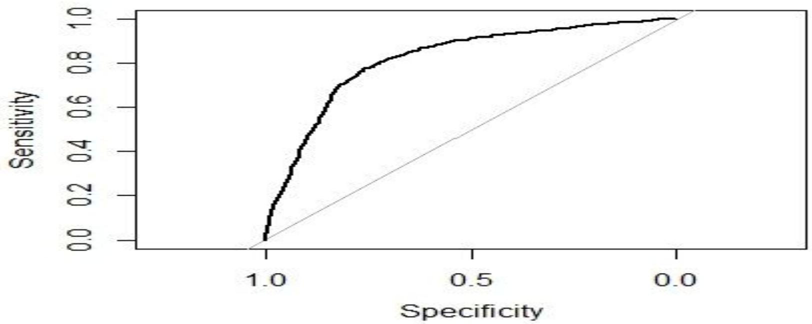
Specificity:TN/(TN+FP))
Logistic regression = 0.317
KNN = 0.571
LDA = 0.475
KNN has a good specificity

F1 score:2*Precision*Recall/(Precision+Recall)
If the data is imbalanced then such kind of a feature is used this calculates accuracy by taking the harmonic mean of precision and recall

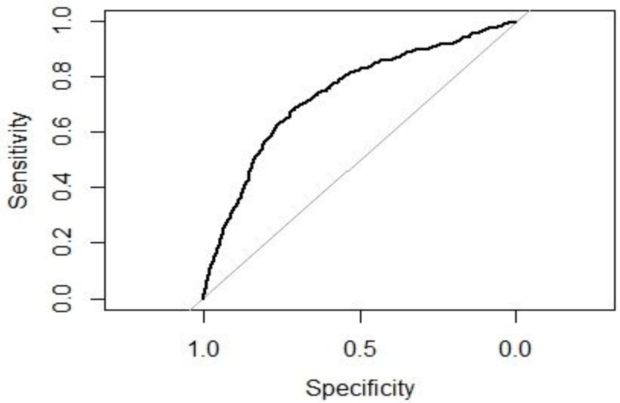
Logistic regression = 0.844
KNN = 0.92
LDA = 0.91

So LDA and KNN are almost pretty accurate so both the models can be considered.

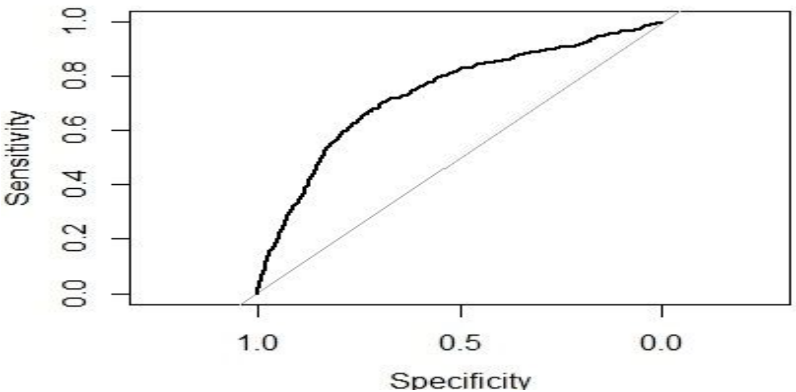
AUC:the area under roc curve
Logistic regression = 0.751
KNN = 0.735
LDA = 0.8177



LDA



Logistic regression



KNN

LDA is the best model in terms of AUC.

Gini Coefficient: (AUC-0.5)/0.5
Logistic regression = 0.5
KNN = 0.47
LDA = 0.635

So LDA is the best model in terms of gini coefficient

Clearly the best model out of all the three models is LDA since while performing different all the performance metrics to the models overall LDA comes out to be the best fit to the data

Actionable insights

Interpretations from the best MODEL

The variables or features considered in this type of model are CustServCalls, contractrenewal,roammins and monthlycharge which are considered in lda which can be a very good predictors of the churn in the future

But in this model clearly there is less percentage of the churn=1 identified successfully so that can improved .that is because of data distribution as the number of customers where churn=0 are very high and for churn=1 are very ,so because of that there are problems.that can be done by considering a probability of (0.5,0.5) for both the target variable and then proceed further which improves the lda output .Since 0.8177 is a good auc value it is definetly a good predictor in the future .