LDA_Customer Churn

14/04/2020

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library (stringr)
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.6.2
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(grid)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.6.2
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.6.2
```

```
## corrplot 0.84 loaded
```

```
library(scales)
library(qqplotr)
```

```
## Warning: package 'qqplotr' was built under R version 3.6.3
```

```
##
## Attaching package: 'qqplotr'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(DMwR)
```

```
## Warning: package 'DMwR' was built under R version 3.6.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 3.6.2
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.3
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.6.3
```

```
library(regclass)
```

```
## Warning: package 'regclass' was built under R version 3.6.3
## Loading required package: bestglm
## Warning: package 'bestglm' was built under R version 3.6.3
## Loading required package: leaps
## Warning: package 'leaps' was built under R version 3.6.3
## Loading required package: VGAM
## Warning: package 'VGAM' was built under R version 3.6.3
## Loading required package: stats4
## Loading required package: splines
##
## Attaching package: 'VGAM'
## The following object is masked from 'package:car':
##
##     logit
## Loading required package: rpart
## Loading required package: randomForest
## Warning: package 'randomForest' was built under R version 3.6.3
## randomForest 4.6-14
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:gridExtra':
##
##     combine
## The following object is masked from 'package:dplyr':
##
##     combine
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## Important regclass change from 1.3:
## All functions that had a . in the name now have an _
## all.correlations -> all_correlations, cor.demo -> cor_demo, etc.

##
## Attaching package: 'regclass'

## The following object is masked from 'package:lattice':
##
##     qq

library(caret)

## Warning: package 'caret' was built under R version 3.6.3

##
## Attaching package: 'caret'

## The following object is masked from 'package:VGAM':
##
##     predictors

library(caTools)

## Warning: package 'caTools' was built under R version 3.6.3

library(pROC)

## Warning: package 'pROC' was built under R version 3.6.3

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.6.3

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.6.3

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages ---------------------------------------------------
------------ tidyverse 1.3.0 --
```

```
## v tibble  2.1.3     v purrr   0.3.3
## v tidyr   1.0.2     v forcats 0.4.0
## v readr   1.3.1
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.2
```

```
## Warning: package 'forcats' was built under R version 3.6.2
```

```
## -- Conflicts ------------------------------------------------------------
------ tidyverse_conflicts() --
## x data.table::between()   masks dplyr::between()
## x readr::col_factor()     masks scales::col_factor()
## x randomForest::combine() masks gridExtra::combine(), dplyr::combine()
## x purrr::discard()        masks scales::discard()
## x tidyr::fill()           masks VGAM::fill()
## x dplyr::filter()         masks stats::filter()
## x data.table::first()     masks dplyr::first()
## x dplyr::lag()            masks stats::lag()
## x data.table::last()      masks dplyr::last()
## x purrr::lift()           masks caret::lift()
## x randomForest::margin()  masks ggplot2::margin()
## x car::recode()           masks dplyr::recode()
## x MASS::select()          masks dplyr::select()
## x purrr::some()           masks car::some()
## x qqplotr::stat_qq_line() masks ggplot2::stat_qq_line()
## x purrr::transpose()      masks data.table::transpose()
```

```
library(MVA)
```

```
## Warning: package 'MVA' was built under R version 3.6.2
```

```
## Loading required package: HSAUR2
```

```
## Warning: package 'HSAUR2' was built under R version 3.6.2
```

```
## Loading required package: tools
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa

library(gvlma)

##-------------------------------------------------------------------------
--------
##Importing Dataset and doing preliminary analysis
##-------------------------------------------------------------------------
--------

#Importing CSV file from drive on my local computer and viewing it

custc <-
read.csv("C:/Users/admin/Desktop/MVA/PROJECT/TelEco_Customer_Churn.csv")
custc <- as.data.frame(custc)
View(custc)

#Checking the Dimension of the dataset

dim(custc)

## [1] 7043   21

#Viewing the first 4 rows of the dataset to get the overview of the dataset

head(custc,4)

##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1 7590-VHVEG Female             0     Yes         No      1           No
## 2 5575-GNVDE   Male             0      No         No     34          Yes
## 3 3668-QPYBK   Male             0      No         No      2          Yes
## 4 7795-CFOCW   Male             0      No         No     45           No
##      MultipleLines InternetService OnlineSecurity OnlineBackup
DeviceProtection
## 1 No phone service             DSL             No          Yes
No
## 2               No             DSL            Yes           No
Yes
## 3               No             DSL            Yes          Yes
No
## 4 No phone service             DSL            Yes           No
Yes
```

```
##   TechSupport StreamingTV StreamingMovies       Contract PaperlessBilling
## 1          No          No              No Month-to-month              Yes
## 2          No          No              No       One year               No
## 3          No          No              No Month-to-month              Yes
## 4         Yes          No              No       One year               No
##               PaymentMethod MonthlyCharges TotalCharges Churn
## 1          Electronic check          29.85        29.85    No
## 2              Mailed check          56.95      1889.50    No
## 3              Mailed check          53.85       108.15   Yes
## 4 Bank transfer (automatic)          42.30      1840.75    No

#Gaining more insight about the kind of data stored in each column

summary(custc)

##       customerID      gender     SeniorCitizen    Partner    Dependents
##  0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933
##  0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110
##  0004-TLHLJ:   1                 Median :0.0000
##  0011-IGKFF:   1                 Mean   :0.1621
##  0013-EXCHZ:   1                 3rd Qu.:0.0000
##  0013-MHZWF:   1                 Max.   :1.0000
##  (Other)   :7037
##      tenure      PhoneService          MultipleLines     InternetService
##  Min.   : 0.00   No : 682     No              :3390   DSL        :2421
##  1st Qu.: 9.00   Yes:6361     No phone service: 682   Fiber optic:3096
##  Median :29.00                Yes             :2971   No         :1526
##  Mean   :32.37
##  3rd Qu.:55.00
##  Max.   :72.00
##
##              OnlineSecurity              OnlineBackup
##  No                 :3498   No                 :3088
##  No internet service:1526   No internet service:1526
##  Yes                :2019   Yes                :2429
##
##
##
##
##             DeviceProtection              TechSupport
##  No                 :3095    No                 :3473
##  No internet service:1526    No internet service:1526
##  Yes                :2422    Yes                :2044
##
##
##
##
##              StreamingTV              StreamingMovies           Contract
##  No                 :2810   No                 :2785   Month-to-month:3875
##  No internet service:1526   No internet service:1526   One year      :1473
```

```
##  Yes                :2707   Yes                :2732   Two year      :1695
##
##
##
##
##  PaperlessBilling                   PaymentMethod  MonthlyCharges
##  No :2872         Bank transfer (automatic):1544   Min.   : 18.25
##  Yes:4171         Credit card (automatic)  :1522   1st Qu.: 35.50
##                   Electronic check         :2365   Median : 70.35
##                   Mailed check             :1612   Mean   : 64.76
##                                                    3rd Qu.: 89.85
##                                                    Max.   :118.75
##
##   TotalCharges    Churn
##  Min.   :  18.8   No :5174
##  1st Qu.: 401.4   Yes:1869
##  Median :1397.5
##  Mean   :2283.3
##  3rd Qu.:3794.7
##  Max.   :8684.8
##  NA's   :11
```

```
glimpse(custc)
```

```
## Observations: 7,043
## Variables: 21
## $ customerID       <fct> 7590-VHVEG, 5575-GNVDE, 3668-QPYBK, 7795-CFOCW,
92...
## $ gender           <fct> Female, Male, Male, Male, Female, Female, Male,
Fe...
## $ SeniorCitizen    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Partner          <fct> Yes, No, No, No, No, No, No, No, Yes, No, Yes,
No,...
## $ Dependents       <fct> No, No, No, No, No, No, Yes, No, No, Yes, Yes,
No,...
## $ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58,
49...
## $ PhoneService     <fct> No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes,
Yes...
## $ MultipleLines    <fct> No phone service, No, No, No phone service, No,
Ye...
## $ InternetService  <fct> DSL, DSL, DSL, DSL, Fiber optic, Fiber optic,
Fibe...
## $ OnlineSecurity   <fct> No, Yes, Yes, Yes, No, No, No, Yes, No, Yes, Yes,
...
## $ OnlineBackup     <fct> Yes, No, Yes, No, No, No, Yes, No, No, Yes, No,
No...
## $ DeviceProtection <fct> No, Yes, No, Yes, No, Yes, No, No, Yes, No, No,
No...
```

```
## $ TechSupport      <fct> No, No, No, Yes, No, No, No, No, Yes, No, No, No
i...
## $ StreamingTV      <fct> No, No, No, No, No, Yes, Yes, No, Yes, No, No, No
...
## $ StreamingMovies  <fct> No, No, No, No, No, Yes, No, No, Yes, No, No, No
i...
## $ Contract         <fct> Month-to-month, One year, Month-to-month, One
year...
## $ PaperlessBilling <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, No,
Yes,...
## $ PaymentMethod    <fct> Electronic check, Mailed check, Mailed check,
Bank...
## $ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10,
2...
## $ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50,
1...
## $ Churn            <fct> No, No, Yes, No, Yes, Yes, No, No, Yes, No, No,
No...

#The above results give us an insight that TotalCharges and MonthlyCharges
are numerical values
#SeniorCitizen and tenure are stored as numerical which need to be converted
to categorical variables

##-------------------------------------------------------------------------
--------
## Performing Data Cleaning and Formatting
##-------------------------------------------------------------------------
--------

#Converting SeniorCitizen numerical variable into Categorical Variable

custc$SeniorCitizen<-factor(custc$SeniorCitizen,levels = c(0 ,1),labels =
c('no','yes'))

#Converting tenure values into ranges of 12 months

custc <- mutate(custc,Tenure_Range =tenure)
cut(custc$Tenure_Range,6,labels = c('0-1 Years','1-2 Years','2-3 Years','4-5
Years','5-6 Years','6-7 Years'))

##    [1] 0-1 Years 2-3 Years 0-1 Years 4-5 Years 0-1 Years 0-1 Years 1-2
Years
##    [8] 0-1 Years 2-3 Years 6-7 Years 1-2 Years 1-2 Years 5-6 Years 5-6
Years
##   [15] 2-3 Years 6-7 Years 5-6 Years 6-7 Years 0-1 Years 1-2 Years 0-1
Years
##   [22] 0-1 Years 0-1 Years 5-6 Years 5-6 Years 2-3 Years 4-5 Years 0-1
Years
##   [29] 6-7 Years 1-2 Years 6-7 Years 0-1 Years 2-3 Years 0-1 Years 0-1
```

```
Years
## [7036] 1-2 Years 0-1 Years 6-7 Years 1-2 Years 6-7 Years 0-1 Years 0-1
Years
## [7043] 6-7 Years
## Levels: 0-1 Years 1-2 Years 2-3 Years 4-5 Years 5-6 Years 6-7 Years
```

```
custc$Tenure_Range <- cut(custc$Tenure_Range,6,labels = c('0-1 Years','1-2
Years','2-3 Years','4-5 Years','5-6 Years','6-7 Years'))

#Checking if there are any NULL values in any of the columns
table(is.na(custc))
```

```
##
##  FALSE   TRUE
## 154935     11
```

```
str_detect(custc,'NA')
```

```
## Warning in stri_detect_regex(string, pattern, negate = negate, opts_regex
=
## opts(pattern)): argument is not an atomic vector; coercing
```

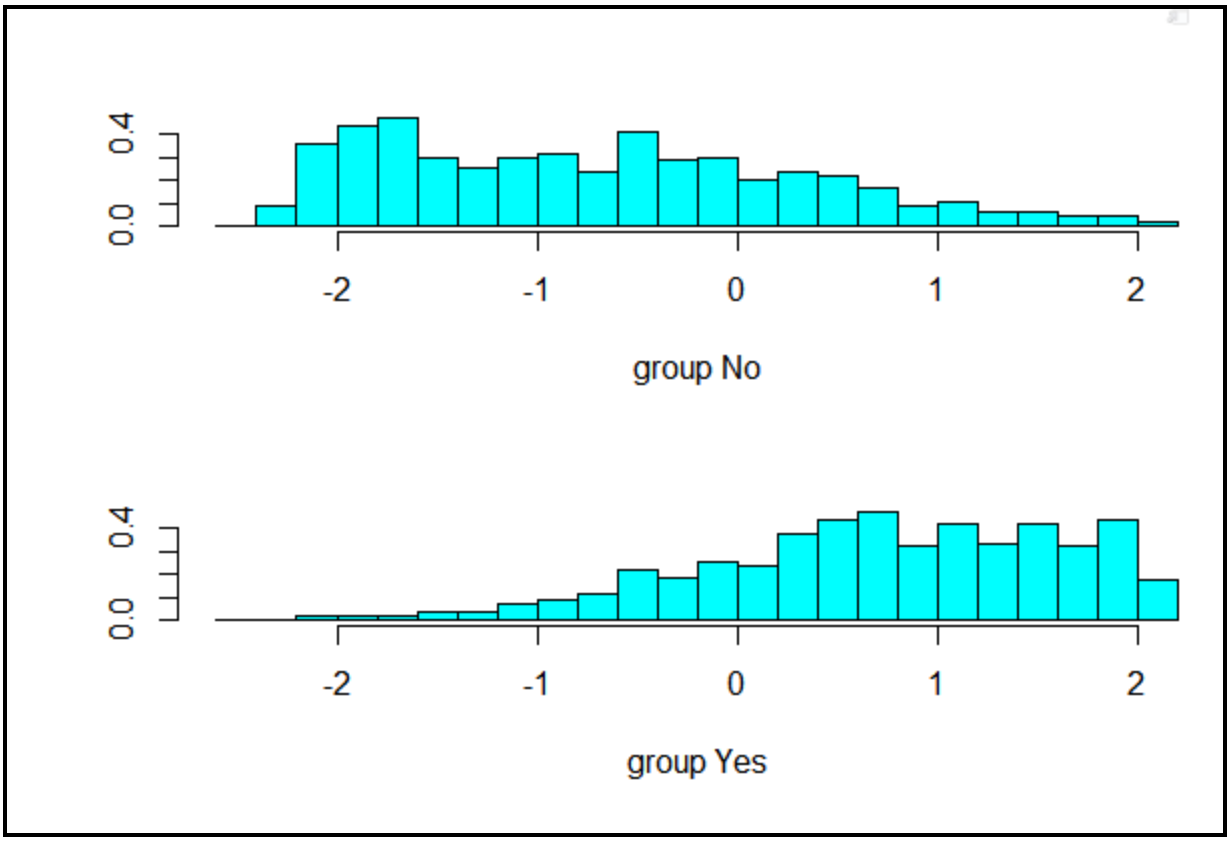
```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
```

```
setDT(custc)
custc[is.na(TotalCharges),NROW(TotalCharges)]
```

```
## [1] 11
```

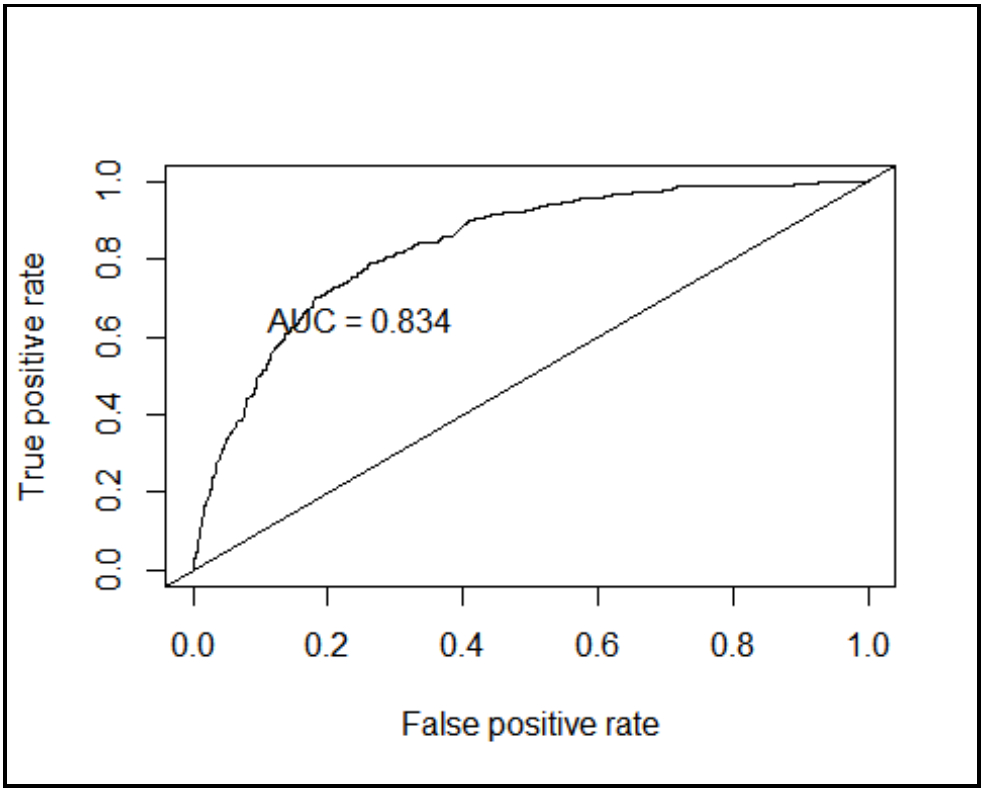
```
#There are 11 rows out of 7043 rows that have null values.Hence removing
these rows since they are only 0.15% of total so we can afford to drop them

custc <- custc[complete.cases(custc), ]

#Replacing 'No Internet Service' values in OnlineSecurity,OnlineBackup
DeviceProtection,TechSupport,StreamingTV and StreamingMovies columns with
'No'

custc$OnlineSecurity[custc$OnlineSecurity=='No internet service'] <- 'No'
custc$OnlineBackup[custc$OnlineBackup=='No internet service'] <- 'No'
custc$DeviceProtection[custc$DeviceProtection=='No internet service'] <- 'No'
custc$TechSupport[custc$TechSupport=='No internet service'] <- 'No'
custc$StreamingTV[custc$StreamingTV=='No internet service'] <- 'No'
custc$StreamingMovies[custc$StreamingMovies=='No internet service'] <- 'No'

#Deleting the unused levels from the factor variables

custc$OnlineSecurity <- factor(custc$OnlineSecurity)
custc$OnlineBackup <- factor(custc$OnlineBackup)
```

```r
custc$DeviceProtection <- factor(custc$DeviceProtection)
custc$TechSupport <- factor(custc$TechSupport)
custc$StreamingTV <- factor(custc$StreamingTV)
custc$StreamingMovies <- factor(custc$StreamingMovies)

##------------------Linear Discriminant Analysis (LDA)---------------------##

##Using same independent variables that we found from logistic regression and
performing LDA to see how well we would be able to predict using this model

custc.data <-(custc[,c("SeniorCitizen","Partner","Dependents","Tenure_Range",

"PhoneService","InternetService","OnlineBackup","OnlineSecurity",
                                 "DeviceProtection","TechSupport","Contract",

"PaperlessBilling","PaymentMethod","Churn")])

##Splitting data into 75% training and 25% test so that we have some data we
can test our model on

smp_size_churn <- floor(0.75 * nrow(custc.data))
train_ind_churn <- sample(nrow(custc.data), size = smp_size_churn)
train_churn.df <- as.data.frame(custc.data[train_ind_churn, ])
test_churn.df <- as.data.frame(custc.data[-train_ind_churn, ])

##Performing LDA on our training data

custc.lda <- lda(Churn~SeniorCitizen+Partner+Dependents+Tenure_Range+

PhoneService+InternetService+OnlineBackup+OnlineSecurity+
                        DeviceProtection+TechSupport+Contract+
                        PaperlessBilling+PaymentMethod, data=train_churn.df)

plot(custc.lda)
```




```
##Making predictions on our testing data

custc.lda.predict <- predict(custc.lda, newdata = test_churn.df)

### CONSTRUCTING ROC AUC PLOT:

# Get the posteriors as a dataframe.
custc.lda.predict.posteriors <- as.data.frame(custc.lda.predict$posterior)
head(custc.lda.predict.posteriors)

##          No        Yes
## 1 0.3698197 0.63018027
## 2 0.7315213 0.26847872
## 3 0.9596578 0.04034223
## 4 0.6616361 0.33836393
## 5 0.5802364 0.41976361
## 6 0.8581338 0.14186621

# Evaluating the model

pred <- prediction(custc.lda.predict.posteriors[,2], test_churn.df$Churn)
roc.perf = performance(pred, measure = "tpr", x.measure = "fpr")
auc.train <- performance(pred, measure = "auc")
auc.train <- auc.train@y.values
```

```r
#Plotting the graph for better visualization

plot(roc.perf)
abline(a=0, b= 1)
text(x = .25, y = .65 ,paste("AUC = ", round(auc.train[[1]],3), sep = ""))
```




```r
##From the above results we see that we get AUC value as 83.5% using LDA
which implies this model is good
##fit and the predictors used in this model can influence our dependent
variable Churn.
```