ChurnLDA.R

2019-11-21

```
##Multivariate Project
##TELECOM-CHURN-ANALYSIS
##Author : Sanket Gohel

##Importing Libraries

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.3

##
## 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)

## Warning: package 'stringr' was built under R version 3.5.3

library(data.table)

## Warning: package 'data.table' was built under R version 3.5.3

##
## 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 3.5.3

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

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

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.5.3
```

```
## corrplot 0.84 loaded
```

```
library(scales)
```

```
## Warning: package 'scales' was built under R version 3.5.3
```

```
library(qqplotr)
```

```
## Warning: package 'qqplotr' was built under R version 3.5.3
```

```
##
## 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 3.5.3
```

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

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.3
```

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

```
## Warning: package 'carData' was built under R version 3.5.2
```

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

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

```
library(e1071)

## Warning: package 'e1071' was built under R version 3.5.3

library(caret)

## Warning: package 'caret' was built under R version 3.5.3

library(caTools)

## Warning: package 'caTools' was built under R version 3.5.3

library(pROC)

## Warning: package 'pROC' was built under R version 3.5.3

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

##
## Attaching package: 'pROC'

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

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.3

## -- Attaching packages ---------------------------------------------------
--- tidyverse 1.2.1 --

## v tibble  2.1.3     v readr   1.3.1
## v tidyr   1.0.0     v purrr   0.3.3
## v tibble  2.1.3     v forcats 0.4.0

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'readr' was built under R version 3.5.3

## Warning: package 'purrr' was built under R version 3.5.3

## Warning: package 'forcats' was built under R version 3.5.3

## -- Conflicts ------------------------------------------------------------ ti
dyverse_conflicts() --
## x data.table::between()   masks dplyr::between()
## x readr::col_factor()     masks scales::col_factor()
## x gridExtra::combine()    masks dplyr::combine()
## x purrr::discard()        masks scales::discard()
## 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 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 3.5.3
```

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

```
## Warning: package 'HSAUR2' was built under R version 3.5.3
```

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

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.5.3
```

```
##
## Attaching package: 'GGally'
```

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

```
library(gvlma)
```

```
## Warning: package 'gvlma' was built under R version 3.5.2
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.5.3
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following objects are masked from 'package:scales':
##
##     alpha, rescale
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 3.5.3

##
## ************************************************************

## Note: As of version 1.0.0, cowplot does not change the

##   default ggplot2 theme anymore. To recover the previous

##   behavior, execute:
##   theme_set(theme_cowplot())

## ************************************************************

library(regclass)

## Warning: package 'regclass' was built under R version 3.5.3

## Loading required package: bestglm

## Warning: package 'bestglm' was built under R version 3.5.3

## Loading required package: leaps

## Warning: package 'leaps' was built under R version 3.5.3

## Loading required package: VGAM

## Warning: package 'VGAM' was built under R version 3.5.3

## Loading required package: stats4

## Loading required package: splines

##
## Attaching package: 'VGAM'

## The following objects are masked from 'package:psych':
##
##     fisherz, logistic, logit

## The following object is masked from 'package:tidyr':
##
##     fill

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

## 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 3.5.3

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:psych':
##
##     outlier

## 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(stats)
library(e1071)
library(pROC)
library(ROCR)

## Warning: package 'ROCR' was built under R version 3.5.3

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.5.3

##
## Attaching package: 'gplots'
```

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

```
##Importing Dataset and doing preliminary analysis

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

tablechurn<-read.csv("C:/Users/SHIVANSHI/Desktop/Priyanshi/MVA/Telecom Churn
Analysis Data.csv")
tablechurn <- as.data.frame(tablechurn)


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

summary(tablechurn)
```

```
##       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
```

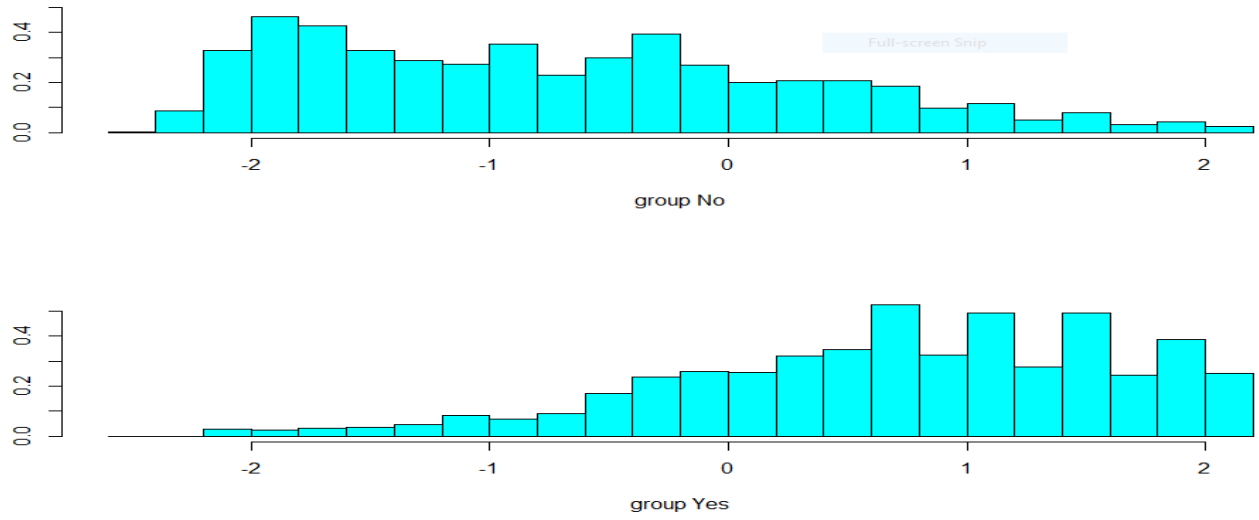
```
glimpse(tablechurn)
```

```
## Observations: 7,043
## Variables: 21
## $ CustomerID       <fct> 7590-VHVEG, 5575-GNVDE, 3668-QPYBK, 7795-CFOCW, 9
2...
## $ Gender           <fct> Female, Male, Male, Male, Female, Female, Male, F
e...
## $ 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, 4
9...
## $ PhoneService     <fct> No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes, Ye
s...
## $ MultipleLines    <fct> No phone service, No, No, No phone service, No, Y
e...
## $ InternetService  <fct> DSL, DSL, DSL, DSL, Fiber optic, Fiber optic, Fib
e...
## $ 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, N
o...
```

```
## $ DeviceProtection <fct> No, Yes, No, Yes, No, Yes, No, No, Yes, No, No, N
o...
## $ 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 yea
r...
## $ PaperlessBilling <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, No, Yes
,...
## $ PaymentMethod    <fct> Electronic check, Mailed check, Mailed check, Ban
k...
## $ 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, N
o...

#The above results give us an insight that TotalCharges and MonthlyCharges ar
e 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

tablechurn$SeniorCitizen<-factor(tablechurn$SeniorCitizen,levels = c(0 ,1),la
bels = c('no','yes'))

#Converting tenure values into ranges of 12 months

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

tablechurn$Tenure_Range <- cut(tablechurn$Tenure_Range,6,labels = c('0-1 Year
s','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(tablechurn))

##
##  FALSE   TRUE
## 154935     11

str_detect(tablechurn,'NA')
```

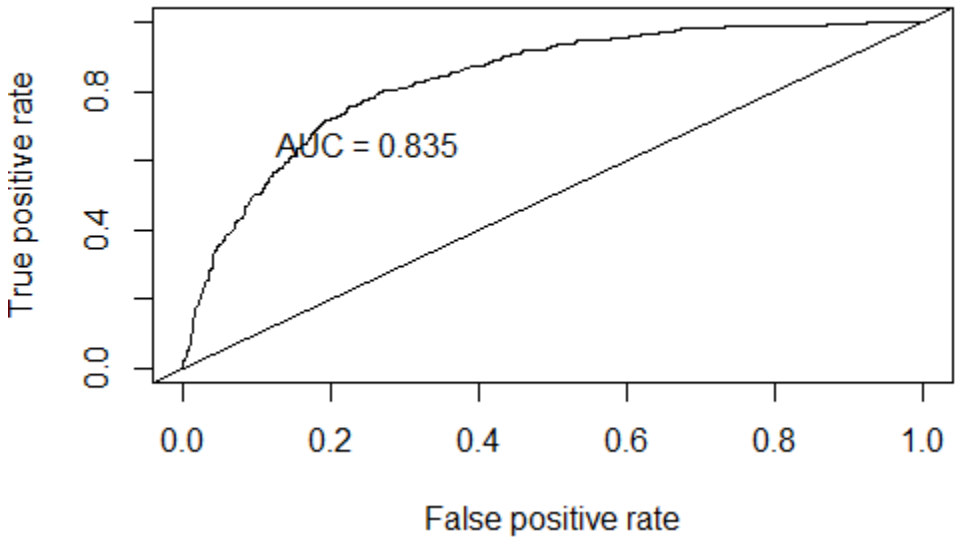
```
setDT(tablechurn)
tablechurn[is.na(TotalCharges),NROW(TotalCharges)]

## [1] 11

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

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

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

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

#Deleting the unused levels from the factor variables

tablechurn$OnlineSecurity <- factor(tablechurn$OnlineSecurity)
tablechurn$OnlineBackup <- factor(tablechurn$OnlineBackup)
tablechurn$DeviceProtection <- factor(tablechurn$DeviceProtection)
tablechurn$TechSupport <- factor(tablechurn$TechSupport)
tablechurn$StreamingTV <- factor(tablechurn$StreamingTV)
tablechurn$StreamingMovies <- factor(tablechurn$StreamingMovies)

##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

tablechurn.data <-(tablechurn[,c("SeniorCitizen","Partner","Dependents","Tenu
re_Range",
                                 "PhoneService","InternetService","OnlineBack
up","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(tablechurn.data))
train_ind_churn <- sample(nrow(tablechurn.data), size = smp_size_churn)
```

```
train_churn.df <- as.data.frame(tablechurn.data[train_ind_churn, ])
test_churn.df <- as.data.frame(tablechurn.data[-train_ind_churn, ])

##Performing LDA on our training data

tablechurn.lda <- lda(Churn~SeniorCitizen+Partner+Dependents+Tenure_Range+
                        PhoneService+InternetService+OnlineBackup+OnlineSecur
ity+
                        DeviceProtection+TechSupport+Contract+
                        PaperlessBilling+PaymentMethod, data=train_churn.df)

plot(tablechurn.lda)
```



```
##Making predictions on our testing data

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

### CONSTRUCTING ROC AUC PLOT:

# Get the posteriors as a dataframe.
tablechurn.lda.predict.posteriors <- as.data.frame(tablechurn.lda.predict$pos
terior)
head(tablechurn.lda.predict.posteriors)

##          No        Yes
## 1 0.4396353 0.56036470
## 2 0.8991660 0.10083402
## 3 0.5377297 0.46227028
## 4 0.9842675 0.01573248
## 5 0.5976701 0.40232988
## 6 0.9713542 0.02864578
```

```r
# Evaluating the model

pred <- prediction(tablechurn.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

#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 whic
h implies this model is good
##fit and the predictors used in this model can influence our dependent varia
ble Churn.
```