Logistic Regression-Customer Churn

09/04/2020

```
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(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
```

```r
##-------------------------------------------------------------------------
--------

#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
##   [36] 6-7 Years 0-1 Years 4-5 Years 2-3 Years 0-1 Years 0-1 Years 6-7
Years
##   [43] 1-2 Years 6-7 Years 1-2 Years 5-6 Years 0-1 Years 0-1 Years 5-6
Years
##   [50] 6-7 Years 4-5 Years 1-2 Years 2-3 Years 0-1 Years 5-6 Years 1-2
Years
##   [57] 6-7 Years 6-7 Years 2-3 Years 6-7 Years 4-5 Years 5-6 Years 6-7
Years
##   [64] 1-2 Years 0-1 Years 0-1 Years 4-5 Years 2-3 Years 5-6 Years 0-1
Years
##   [71] 0-1 Years 5-6 Years 6-7 Years 6-7 Years 0-1 Years 5-6 Years 4-5
Years
##   [78] 0-1 Years 2-3 Years 4-5 Years 0-1 Years 0-1 Years 0-1 Years 4-5
Years
```

```
Years
## [6910] 0-1 Years 6-7 Years 5-6 Years 0-1 Years 6-7 Years 4-5 Years 6-7
Years
## [6917] 4-5 Years 6-7 Years 2-3 Years 5-6 Years 2-3 Years 5-6 Years 0-1
Years
## [6924] 5-6 Years 0-1 Years 1-2 Years 5-6 Years 0-1 Years 4-5 Years 2-3
Years
## [6931] 0-1 Years 5-6 Years 0-1 Years 0-1 Years 6-7 Years 6-7 Years 0-1
Years
## [6938] 2-3 Years 6-7 Years 2-3 Years 6-7 Years 6-7 Years 6-7 Years 0-1
Years
## [6945] 0-1 Years 6-7 Years 4-5 Years 6-7 Years 4-5 Years 2-3 Years 0-1
Years
## [6952] 5-6 Years 4-5 Years 1-2 Years 1-2 Years 0-1 Years 6-7 Years 0-1
Years
## [6959] 1-2 Years 4-5 Years 4-5 Years 1-2 Years 2-3 Years 0-1 Years 5-6
Years
## [6966] 6-7 Years 5-6 Years 2-3 Years 1-2 Years 0-1 Years 0-1 Years 1-2
Years
## [6973] 5-6 Years 5-6 Years 5-6 Years 1-2 Years 6-7 Years 1-2 Years 6-7
Years
## [6980] 0-1 Years 1-2 Years 0-1 Years 6-7 Years 1-2 Years 2-3 Years 4-5
Years
## [6987] 2-3 Years 2-3 Years 1-2 Years 1-2 Years 2-3 Years 0-1 Years 6-7
Years
## [6994] 5-6 Years 4-5 Years 5-6 Years 4-5 Years 2-3 Years 1-2 Years 0-1
Years
## [7001] 6-7 Years 0-1 Years 6-7 Years 2-3 Years 4-5 Years 1-2 Years 4-5
Years
## [7008] 6-7 Years 0-1 Years 1-2 Years 0-1 Years 0-1 Years 6-7 Years 4-5
Years
## [7015] 4-5 Years 2-3 Years 0-1 Years 5-6 Years 0-1 Years 4-5 Years 0-1
Years
## [7022] 0-1 Years 6-7 Years 6-7 Years 4-5 Years 1-2 Years 0-1 Years 1-2
Years
## [7029] 6-7 Years 0-1 Years 0-1 Years 5-6 Years 0-1 Years 4-5 Years 6-7
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)
custc$DeviceProtection <- factor(custc$DeviceProtection)
custc$TechSupport <- factor(custc$TechSupport)
custc$StreamingTV <- factor(custc$StreamingTV)
custc$StreamingMovies <- factor(custc$StreamingMovies)

##----------------LOGISTIC REGRESSION----------------------------------------
----##

##Checking relationships between our dependent variable and each of our
independent categorical variable.

xtabs(~Churn+gender,data=custc)
```

```
##      gender
## Churn Female Male
##   No    2544 2619
##   Yes    939  930
```

```
xtabs(~Churn+SeniorCitizen,data=custc)
```

```
##      SeniorCitizen
## Churn   no  yes
##   No  4497  666
##   Yes 1393  476
```

```
xtabs(~Churn+Partner,data=custc)
```

```
##      Partner
## Churn   No  Yes
##   No  2439 2724
##   Yes 1200  669
```

```
xtabs(~Churn+Dependents,data=custc)
```

```
##      Dependents
## Churn   No  Yes
##   No  3390 1773
##   Yes 1543  326
```

```
xtabs(~Churn+Tenure_Range,data=custc)
```

```
##      Tenure_Range
## Churn 0-1 Years 1-2 Years 2-3 Years 4-5 Years 5-6 Years 6-7 Years
##   No       1138       730       652       617       712      1314
##   Yes      1037       294       180       145       120        93
```

```
xtabs(~Churn+PhoneService,data=custc)
```

```
##      PhoneService
## Churn   No  Yes
##   No   510 4653
##   Yes  170 1699
```

```
xtabs(~Churn+MultipleLines,data=custc)
```

```
##      MultipleLines
## Churn   No No phone service  Yes
##   No  2536              510 2117
##   Yes  849              170  850
```

```
xtabs(~Churn+InternetService,data=custc)
```

```
##      InternetService
## Churn  DSL Fiber optic   No
##   No  1957        1799 1407
##   Yes  459        1297  113
```

```
xtabs(~Churn+OnlineBackup,data=custc)
```

```
##      OnlineBackup
## Churn   No  Yes
##   No  3261 1902
##   Yes 1346  523
```

```
xtabs(~Churn+OnlineSecurity,data=custc)
```

```
##      OnlineSecurity
## Churn   No  Yes
##   No  3443 1720
##   Yes 1574  295
```

```
xtabs(~Churn+DeviceProtection,data=custc)
```

```
##      DeviceProtection
## Churn   No  Yes
##   No  3290 1873
##   Yes 1324  545
```

```
xtabs(~Churn+TechSupport,data=custc)
```

```
##      TechSupport
## Churn   No  Yes
##   No  3433 1730
##   Yes 1559  310
```

```
xtabs(~Churn+StreamingTV,data=custc)
```

```
##      StreamingTV
## Churn   No  Yes
##   No  3274 1889
##   Yes 1055  814
```

```
xtabs(~Churn+StreamingMovies,data=custc)
```

```
##      StreamingMovies
## Churn   No  Yes
##   No  3250 1913
##   Yes 1051  818
```

```
xtabs(~Churn+Contract,data=custc)
```

```
##      Contract
## Churn Month-to-month One year Two year
##   No            2220     1306     1637
##   Yes           1655      166       48
```

```
xtabs(~Churn+PaperlessBilling,data=custc)
```

```
##      PaperlessBilling
## Churn   No  Yes
```

```
##   No  2395 2768
##   Yes  469 1400
```

```
xtabs(~Churn+PaymentMethod,data=custc)
```

```
##      PaymentMethod
## Churn Bank transfer (automatic) Credit card (automatic) Electronic check
##   No                      1284                    1289              1294
##   Yes                      258                     232              1071
##      PaymentMethod
## Churn Mailed check
##   No          1296
##   Yes          308
```

```
##By above results, we find that independent variables like Senior Citizen
,Partner, Dependents,
##Tenure Range,Phone Service,Internet
Service,OnlineBackup,OnlineSecurity,DeviceProtection,
##TechSupport,StreamingTV,StreamingMovies,Contract,PaperLess Billing,Payment
Method variables can
##have impact on dependent variable (Churn).
##Although we see that the variables like StreamingTV and StreamingMovies
don't show significant difference in
##indicating if person will churn or not based on the result.
##So lets run 2 model.One simple model excluding StreamingTV and
StreamingMovies and other
##including all independent variables mentioned above.
```

```
logistic_simple <- glm(Churn~SeniorCitizen+Partner+Dependents+Tenure_Range+

PhoneService+InternetService+OnlineBackup+OnlineSecurity+
                       DeviceProtection+TechSupport+Contract+
                       PaperlessBilling+PaymentMethod, data=custc,
family="binomial")
summary(logistic_simple)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Partner + Dependents +
##     Tenure_Range + PhoneService + InternetService + OnlineBackup +
##     OnlineSecurity + DeviceProtection + TechSupport + Contract +
##     PaperlessBilling + PaymentMethod, family = "binomial", data = custc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8299  -0.6813  -0.2962   0.7144   3.0380
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.18898    0.14636  -1.291 0.196623
```

```
## SeniorCitizenyes                          0.23893    0.08391   2.847 0.004410
**
## PartnerYes                               -0.01211    0.07720  -0.157 0.875307
## DependentsYes                            -0.14288    0.08924  -1.601 0.109351
## Tenure_Range1-2 Years                    -0.82050    0.09518  -8.621  < 2e-16
***
## Tenure_Range2-3 Years                    -1.15770    0.11255 -10.286  < 2e-16
***
## Tenure_Range4-5 Years                    -1.07778    0.12593  -8.559  < 2e-16
***
## Tenure_Range5-6 Years                    -1.29415    0.13682  -9.459  < 2e-16
***
## Tenure_Range6-7 Years                    -1.50332    0.16432  -9.149  < 2e-16
***
## PhoneServiceYes                          -0.43054    0.12379  -3.478 0.000505
***
## InternetServiceFiber optic                1.04119    0.09013  11.552  < 2e-16
***
## InternetServiceNo                        -0.93730    0.13656  -6.864 6.71e-12
***
## OnlineBackupYes                          -0.15218    0.07579  -2.008 0.044647
*
## OnlineSecurityYes                        -0.40059    0.08386  -4.777 1.78e-06
***
## DeviceProtectionYes                       0.06687    0.07671   0.872 0.383348
## TechSupportYes                           -0.28669    0.08438  -3.397 0.000680
***
## ContractOne year                         -0.73328    0.10626  -6.901 5.17e-12
***
## ContractTwo year                         -1.61097    0.18056  -8.922  < 2e-16
***
## PaperlessBillingYes                       0.37588    0.07394   5.084 3.70e-07
***
## PaymentMethodCredit card (automatic)     -0.08137    0.11301  -0.720 0.471521
## PaymentMethodElectronic check             0.36328    0.09358   3.882 0.000104
***
## PaymentMethodMailed check                -0.02452    0.11367  -0.216 0.829239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8143.4  on 7031  degrees of freedom
## Residual deviance: 5896.7  on 7010  degrees of freedom
## AIC: 5940.7
##
## Number of Fisher Scoring iterations: 6

## Calculating the p-value for R^2 for this model
```

```r
ll.null <- logistic_simple$null.deviance/-2
ll.proposed <- logistic_simple$deviance/-2
(ll.null - ll.proposed) / ll.null
```

```
## [1] 0.275893
```

```r
1 - pchisq(2*(ll.proposed - ll.null),
df=(length(logistic_simple$coefficients)-1))
```

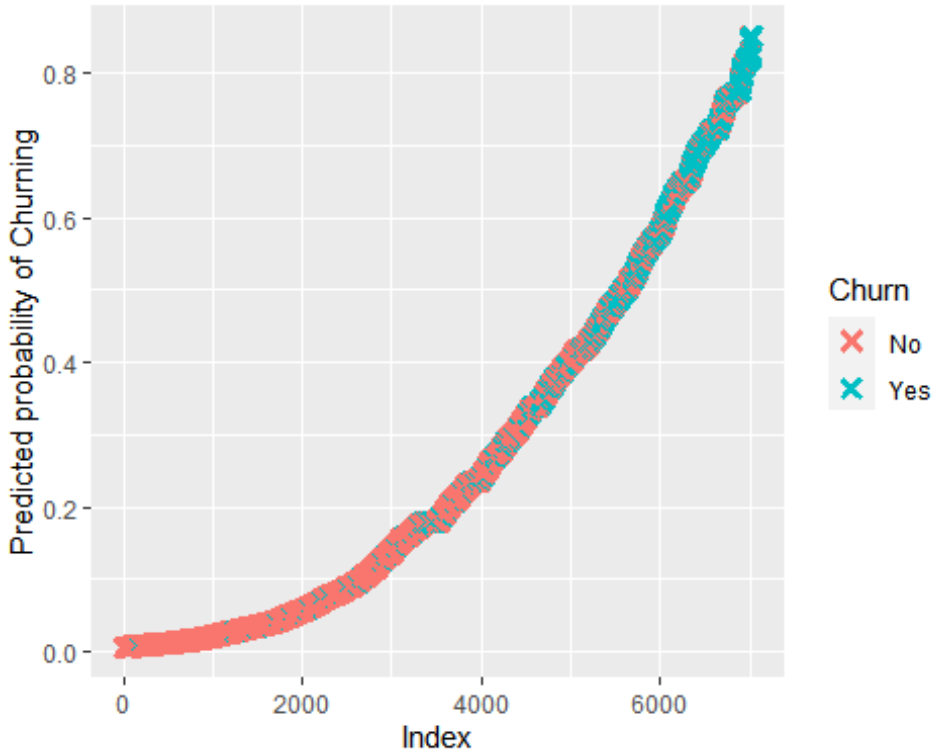
```
## [1] 0
```

```r
##Performing regression using all variables including StreamingTV and
StreamingMovies

logistic <- glm(Churn~SeniorCitizen+Partner+Dependents+Tenure_Range+
                  PhoneService+InternetService+OnlineBackup+OnlineSecurity+

DeviceProtection+TechSupport+StreamingTV+StreamingMovies+Contract+
                  PaperlessBilling+PaymentMethod,data=custc,
family="binomial")
summary(logistic)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Partner + Dependents +
##     Tenure_Range + PhoneService + InternetService + OnlineBackup +
##     OnlineSecurity + DeviceProtection + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod,
##     family = "binomial", data = custc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9509  -0.6761  -0.2862   0.6852   3.1013
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.268442   0.147838  -1.816 0.069404
.
## SeniorCitizenyes                 0.233841   0.084121   2.780 0.005439
**
## PartnerYes                      -0.021832   0.077456  -0.282 0.778053
## DependentsYes                   -0.135752   0.089529  -1.516 0.129446
## Tenure_Range1-2 Years           -0.868833   0.096153  -9.036  < 2e-16
***
## Tenure_Range2-3 Years           -1.229227   0.114174 -10.766  < 2e-16
***
## Tenure_Range4-5 Years           -1.161742   0.127917  -9.082  < 2e-16
***
## Tenure_Range5-6 Years           -1.409129   0.139231 -10.121  < 2e-16
***
## Tenure_Range6-7 Years           -1.629799   0.166836  -9.769  < 2e-16
```

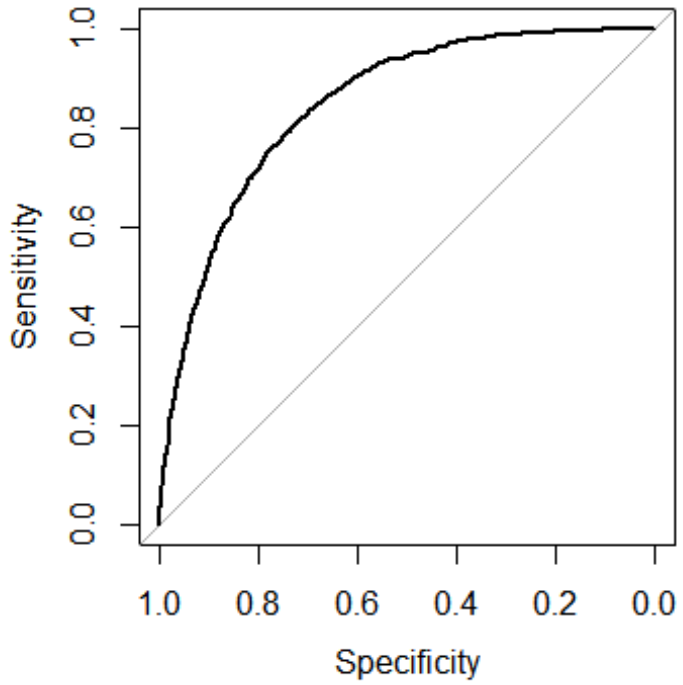
```
***
## PhoneServiceYes                          -0.392387   0.124612  -3.149 0.001639
**
## InternetServiceFiber optic                0.953028   0.091513  10.414  < 2e-16
***
## InternetServiceNo                        -0.851492   0.137644  -6.186 6.16e-10
***
## OnlineBackupYes                          -0.157162   0.076153  -2.064 0.039040
*
## OnlineSecurityYes                        -0.384573   0.084121  -4.572 4.84e-06
***
## DeviceProtectionYes                      -0.020375   0.078462  -0.260 0.795107
## TechSupportYes                           -0.334370   0.085226  -3.923 8.73e-05
***
## StreamingTVYes                            0.271152   0.079621   3.406 0.000660
***
## StreamingMoviesYes                        0.282140   0.079520   3.548 0.000388
***
## ContractOne year                         -0.793518   0.107181  -7.404 1.33e-13
***
## ContractTwo year                         -1.673931   0.181449  -9.225  < 2e-16
***
## PaperlessBillingYes                       0.336435   0.074525   4.514 6.35e-06
***
## PaymentMethodCredit card (automatic)     -0.075260   0.113446  -0.663 0.507076
## PaymentMethodElectronic check             0.325815   0.094142   3.461 0.000538
***
## PaymentMethodMailed check                -0.004958   0.114123  -0.043 0.965346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8143.4  on 7031  degrees of freedom
## Residual deviance: 5861.7  on 7008  degrees of freedom
## AIC: 5909.7
##
## Number of Fisher Scoring iterations: 6

##Calculating p value for R^2 for this model

ll.null <- logistic$null.deviance/-2
ll.proposed <- logistic$deviance/-2
(ll.null - ll.proposed) / ll.null

## [1] 0.2801921

1 - pchisq(2*(ll.proposed - ll.null), df=(length(logistic$coefficients)-1))

## [1] 0
```

```
##Plotting the graphs to visually view this regression

predicted.data <-
data.frame(probability.of.Churn=logistic$fitted.values,Churn=custc$Churn)
predicted.data <- predicted.data[order(predicted.data$probability.of.Churn,
decreasing=FALSE),]
predicted.data$rank <- 1:nrow(predicted.data)

##Plotting the predicted probabilities for each samples probability of
Churning and using colors to visually analyze if they Churned or not

ggplot(data=predicted.data,aes(x=rank, y=probability.of.Churn)) +
  geom_point(aes(color=Churn), alpha=1, shape=4, stroke=2) +
  xlab("Index") +ylab("Predicted probability of Churning")
```



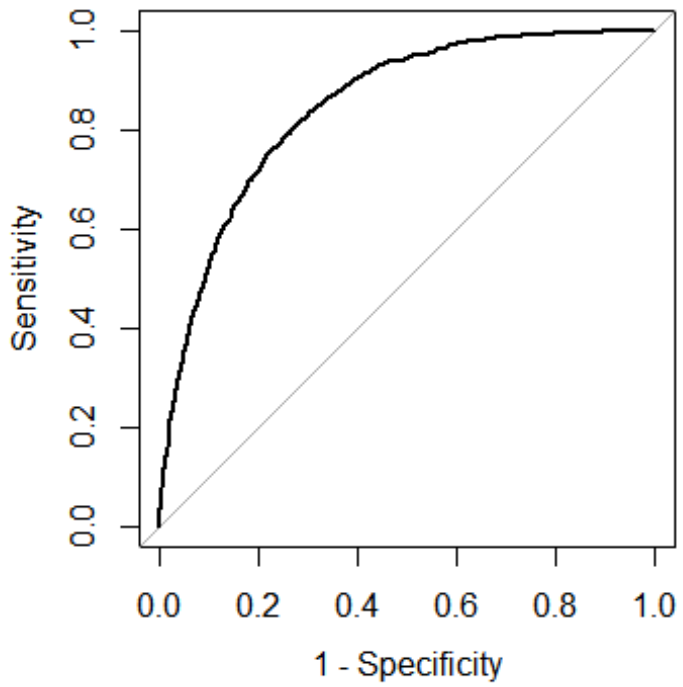

```
##Viewing the confusion matrix for this model

confusion_matrix(logistic)

##            Predicted No Predicted Yes Total
## Actual No          4674           489  5163
## Actual Yes          918           951  1869
## Total              5592          1440  7032

##ROC graph is a plot of the true positive rate against the false positive
rate.Hence plotting it for visualization
```

```
roc(custc$Churn,logistic$fitted.values,plot=TRUE)

## Setting levels: control = No, case = Yes

## Setting direction: controls < cases

##
## Call:
## roc.default(response = custc$Churn, predictor = logistic$fitted.values,
plot = TRUE)
##
## Data: logistic$fitted.values in 5163 controls (custc$Churn No) < 1869
cases (custc$Churn Yes).
## Area under the curve: 0.8454

par(pty='s')
roc(custc$Churn,logistic$fitted.values,plot=TRUE)

## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
```

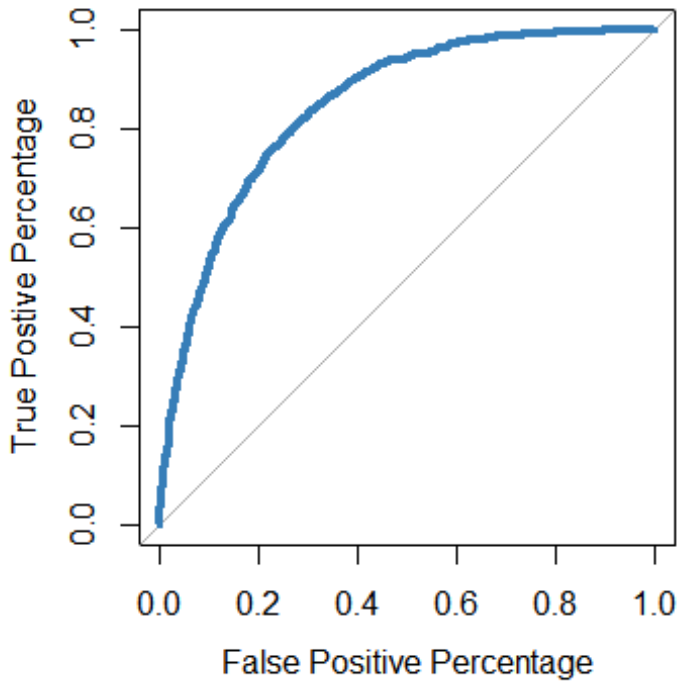



```
##
## Call:
## roc.default(response = custc$Churn, predictor = logistic$fitted.values,
plot = TRUE)
##
## Data: logistic$fitted.values in 5163 controls (custc$Churn No) < 1869
```

```
cases (custc$Churn Yes).
## Area under the curve: 0.8454

##Using 1-specificity (i.e. the False Positive Rate) on the x-axis by setting
"legacy.axes" to TRUE for better visual analysis

roc(custc$Churn,logistic$fitted.values,plot=TRUE, legacy.axes=TRUE)

## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = custc$Churn, predictor = logistic$fitted.values,
plot = TRUE, legacy.axes = TRUE)
##
## Data: logistic$fitted.values in 5163 controls (custc$Churn No) < 1869
cases (custc$Churn Yes).
## Area under the curve: 0.8454

roc(custc$Churn,logistic$fitted.values,plot=TRUE, legacy.axes=TRUE,
xlab="False Positive Percentage", ylab="True Postive Percentage",
col="#377eb8", lwd=4)

## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = custc$Churn, predictor = logistic$fitted.values,
plot = TRUE, legacy.axes = TRUE, xlab = "False Positive Percentage",     ylab
= "True Postive Percentage", col = "#377eb8", lwd = 4)
##
## Data: logistic$fitted.values in 5163 controls (custc$Churn No) < 1869
cases (custc$Churn Yes).
## Area under the curve: 0.8454

## If we want to find out the optimal threshold we can store the data used to
make the ROC graph in a variable

roc.info <- roc(custc$Churn, logistic$fitted.values, legacy.axes=TRUE)

## Setting levels: control = No, case = Yes
## Setting direction: controls < cases

str(roc.info)

## List of 15
##  $ percent           : logi FALSE
##  $ sensitivities     : num [1:4369] 1 1 1 1 1 1 1 1 1 1 ...
##  $ specificities     : num [1:4369] 0 0.000581 0.000775 0.005617 0.006198
...
##  $ thresholds        : num [1:4369] -Inf 0.00616 0.0063 0.00645 0.00653
...
```

```
##  $ direction         : chr "<"
##  $ cases             : Named num [1:1869] 0.295 0.722 0.816 0.476 0.4 ...
##   ..- attr(*, "names")= chr [1:1869] "3" "5" "6" "9" ...
##  $ controls          : Named num [1:5163] 0.5535 0.0434 0.0491 0.4165
0.3412 ...
##   ..- attr(*, "names")= chr [1:5163] "1" "2" "4" "7" ...
##  $ fun.sesp          :function (thresholds, controls, cases, direction)
##  $ auc               : 'auc' num 0.845
##   ..- attr(*, "partial.auc")= logi FALSE
##   ..- attr(*, "percent")= logi FALSE
##   ..- attr(*, "roc")=List of 15
##   .. ..$ percent           : logi FALSE
##   .. ..$ sensitivities     : num [1:4369] 1 1 1 1 1 1 1 1 1 1 ...
##   .. ..$ specificities     : num [1:4369] 0 0.000581 0.000775 0.005617
0.006198 ...
##   .. ..$ thresholds        : num [1:4369] -Inf 0.00616 0.0063 0.00645
0.00653 ...
##   .. ..$ direction         : chr "<"
##   .. ..$ cases             : Named num [1:1869] 0.295 0.722 0.816 0.476
0.4 ...
##   .. .. ..- attr(*, "names")= chr [1:1869] "3" "5" "6" "9" ...
##   .. ..$ controls          : Named num [1:5163] 0.5535 0.0434 0.0491
0.4165 0.3412 ...
##   .. .. ..- attr(*, "names")= chr [1:5163] "1" "2" "4" "7" ...
##   .. ..$ fun.sesp          :function (thresholds, controls, cases,
direction)
##   .. ..$ auc               : 'auc' num 0.845
##   .. .. ..- attr(*, "partial.auc")= logi FALSE
##   .. .. ..- attr(*, "percent")= logi FALSE
##   .. .. ..- attr(*, "roc")=List of 8
##   .. .. .. ..$ percent      : logi FALSE
##   .. .. .. ..$ sensitivities: num [1:4369] 1 1 1 1 1 1 1 1 1 1 ...
##   .. .. .. ..$ specificities: num [1:4369] 0 0.000581 0.000775 0.005617
0.006198 ...
##   .. .. .. ..$ thresholds   : num [1:4369] -Inf 0.00616 0.0063 0.00645
0.00653 ...
##   .. .. .. ..$ direction    : chr "<"
##   .. .. .. ..$ cases        : Named num [1:1869] 0.295 0.722 0.816 0.476
0.4 ...
##   .. .. .. .. ..- attr(*, "names")= chr [1:1869] "3" "5" "6" "9" ...
##   .. .. .. ..$ controls     : Named num [1:5163] 0.5535 0.0434 0.0491
0.4165 0.3412 ...
##   .. .. .. .. ..- attr(*, "names")= chr [1:5163] "1" "2" "4" "7" ...
##   .. .. .. ..$ fun.sesp     :function (thresholds, controls, cases,
direction)
##   .. .. .. ..- attr(*, "class")= chr "roc"
##   .. ..$ call              : language roc.default(response = custc$Churn,
predictor = logistic$fitted.values,      legacy.axes = TRUE)
##   .. ..$ original.predictor: Named num [1:7032] 0.5535 0.0434 0.295 0.0491
0.722 ...
```

```
##   .. .. ..- attr(*, "names")= chr [1:7032] "1" "2" "3" "4" ...
##   .. ..$ original.response : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1
1 2 1 ...
##   .. ..$ predictor         : Named num [1:7032] 0.5535 0.0434 0.295 0.0491
0.722 ...
##   .. .. ..- attr(*, "names")= chr [1:7032] "1" "2" "3" "4" ...
##   .. ..$ response          : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1
1 2 1 ...
##   .. ..$ levels            : chr [1:2] "No" "Yes"
##   .. ..- attr(*, "class")= chr "roc"
##  $ call              : language roc.default(response = custc$Churn,
predictor = logistic$fitted.values,      legacy.axes = TRUE)
##  $ original.predictor: Named num [1:7032] 0.5535 0.0434 0.295 0.0491 0.722
...
##   ..- attr(*, "names")= chr [1:7032] "1" "2" "3" "4" ...
##  $ original.response : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1
...
##  $ predictor         : Named num [1:7032] 0.5535 0.0434 0.295 0.0491 0.722
...
##   ..- attr(*, "names")= chr [1:7032] "1" "2" "3" "4" ...
##  $ response          : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1
...
##  $ levels            : chr [1:2] "No" "Yes"
##  - attr(*, "class")= chr "roc"
```

```
roc.df <- data.frame(tpp=roc.info$sensitivities*100, ## tpp = true positive
percentage
                     fpp=(1 - roc.info$specificities)*100, ## fpp = false
positive precentage
                     thresholds=roc.info$thresholds)

##This will show us the values for the upper right-hand corner of the ROC
graph, when the threshold is so low

head(roc.df)
```

```
##   tpp       fpp  thresholds
## 1 100 100.00000        -Inf
## 2 100  99.94189 0.006159113
## 3 100  99.92253 0.006298493
## 4 100  99.43831 0.006445415
## 5 100  99.38021 0.006526601
## 6 100  99.32210 0.006554029
```

```
##This will show us the values for the lower left-hand corner of the ROC
graph, when the threshold is so high (infinity)

tail(roc.df)
```

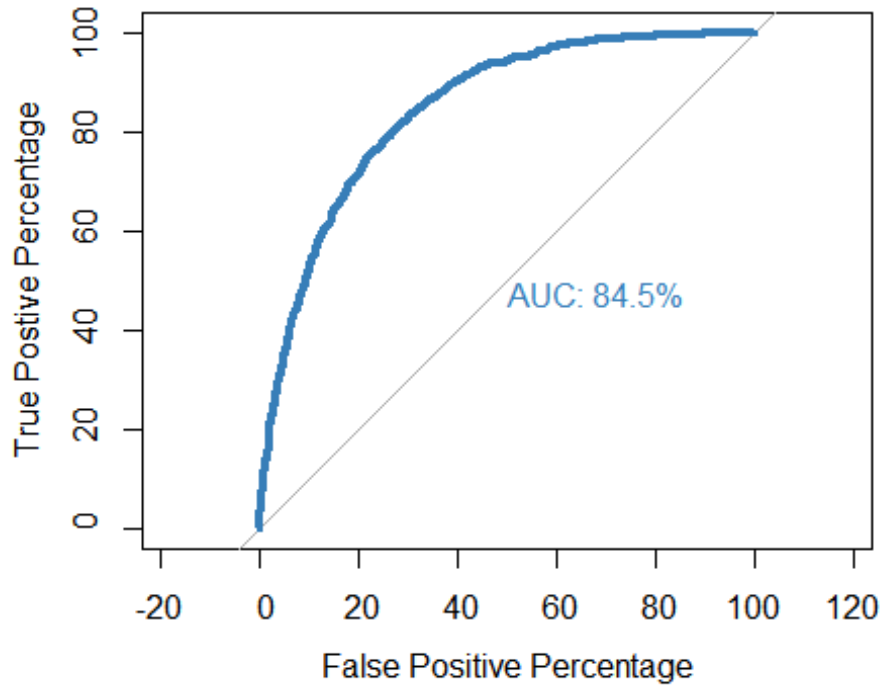
```
##            tpp        fpp thresholds
## 4364 1.3911182 0.09684292  0.8283811
```

```
## 4365 1.2841091 0.09684292  0.8376412
## 4366 1.0165864 0.09684292  0.8467748
## 4367 0.8560728 0.07747434  0.8481903
## 4368 0.6420546 0.05810575  0.8495859
## 4369 0.0000000 0.00000000        Inf
```

```
##Viewing graphs using percentage values
roc(custc$Churn,logistic$fitted.values,plot=TRUE, legacy.axes=TRUE,
xlab="False Positive Percentage", ylab="True Postive Percentage",
col="#377eb8", lwd=4, percent=TRUE, print.auc=TRUE)
```

```
## Setting levels: control = No, case = Yes
```

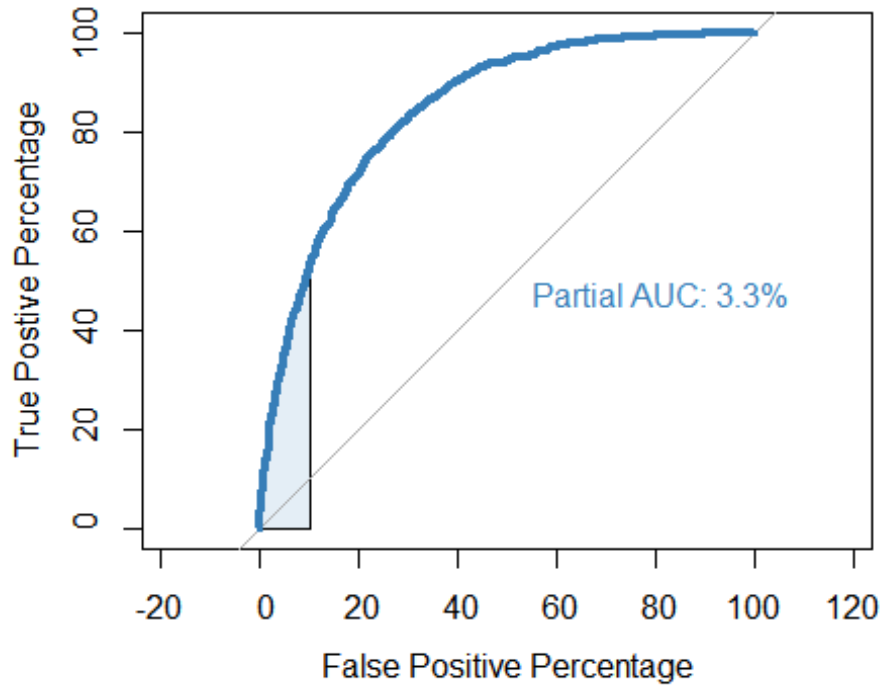
```
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = custc$Churn, predictor = logistic$fitted.values,
percent = TRUE, plot = TRUE, legacy.axes = TRUE, xlab = "False Positive
Percentage",     ylab = "True Postive Percentage", col = "#377eb8", lwd = 4,
print.auc = TRUE)
##
## Data: logistic$fitted.values in 5163 controls (custc$Churn No) < 1869
cases (custc$Churn Yes).
## Area under the curve: 84.54%
```

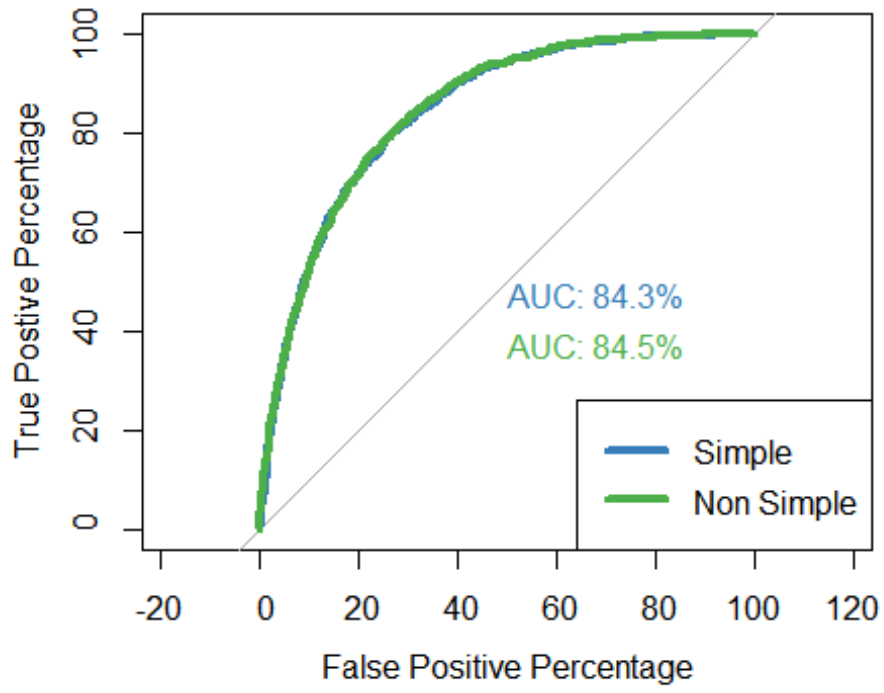
```
roc(custc$Churn,logistic$fitted.values,plot=TRUE, legacy.axes=TRUE,
xlab="False Positive Percentage", ylab="True Postive Percentage",
```

```
col="#377eb8", lwd=4, percent=TRUE, print.auc=TRUE, partial.auc=c(100, 90),
auc.polygon = TRUE, auc.polygon.col = "#377eb822", print.auc.x=45)

## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = custc$Churn, predictor = logistic$fitted.values,
percent = TRUE, plot = TRUE, legacy.axes = TRUE, xlab = "False Positive
Percentage",     ylab = "True Postive Percentage", col = "#377eb8", lwd = 4,
print.auc = TRUE, partial.auc = c(100, 90), auc.polygon = TRUE,
auc.polygon.col = "#377eb822", print.auc.x = 45)
##
## Data: logistic$fitted.values in 5163 controls (custc$Churn No) < 1869
cases (custc$Churn Yes).
## Partial area under the curve (specificity 100%-90%): 3.256%

# Lets do two ROC plots to understand which model is better
roc(custc$Churn, logistic_simple$fitted.values, plot=TRUE, legacy.axes=TRUE,
percent=TRUE, xlab="False Positive Percentage", ylab="True Postive
Percentage", col="#377eb8", lwd=4, print.auc=TRUE)

## Setting levels: control = No, case = Yes

## Setting direction: controls < cases
```

```
##
## Call:
## roc.default(response = custc$Churn, predictor =
logistic_simple$fitted.values,     percent = TRUE, plot = TRUE, legacy.axes =
TRUE, xlab = "False Positive Percentage",     ylab = "True Postive
Percentage", col = "#377eb8", lwd = 4,     print.auc = TRUE)
##
## Data: logistic_simple$fitted.values in 5163 controls (custc$Churn No) <
1869 cases (custc$Churn Yes).
## Area under the curve: 84.34%

# Lets add the other graph
plot.roc(custc$Churn, logistic$fitted.values, percent=TRUE, col="#4daf4a",
lwd=4, print.auc=TRUE, add=TRUE, print.auc.y=40)

## Setting levels: control = No, case = Yes
## Setting direction: controls < cases

legend("bottomright", legend=c("Simple", "Non Simple"), col=c("#377eb8",
"#4daf4a"), lwd=4) # Make it user friendly
```




```
# reset the par area back to the default setting
par(pty='m')
##From the above results we see that we get AUC value as 84.5% with the
second model(i.e. non simple model) which implies this model is good
##fit and the predictors used in this model can influence our dependent
variable Churn.
```