Binary Logistic Regression For Predicting identifying if a customer is churned or not.

```
data<- read.csv("C:/Users/mvpra/OneDrive/Desktop/prac2/WA_Fn-UseC_-Telco-Customer-Churn.csv")
head(data)
```

```
##   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
## 5 9237-HQITU Female             0      No         No      2          Yes
## 6 9305-CDSKC Female             0      No         No      8          Yes
##      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
## 5               No     Fiber optic             No           No
No
## 6              Yes     Fiber optic             No           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
## 5          No          No              No Month-to-month              Yes
## 6          No         Yes             Yes Month-to-month              Yes
##                PaymentMethod MonthlyCharges TotalCharges Churn Churn_int
## 1           Electronic check          29.85        29.85    No         0
## 2               Mailed check          56.95      1889.50    No         0
## 3               Mailed check          53.85       108.15   Yes         1
## 4 Bank transfer (automatic)          42.30      1840.75    No         0
## 5           Electronic check          70.70       151.65   Yes         1
## 6           Electronic check          99.65       820.50   Yes         1
```

```
str(data)
```

```
## 'data.frame':    7043 obs. of  22 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-
```

```
CFOCW" ...
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service"
...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month"
"One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check"
"Bank transfer (automatic)" ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
##  $ Churn_int       : int  0 0 1 0 1 1 0 0 1 0 ...

dim(data)

## [1] 7043   22

sum(which(data$MonthlyCharges>data$TotalCharges))#inconsistencies not present

## [1] 0

summary(data)

##   customerID          gender          SeniorCitizen      Partner
##  Length:7043        Length:7043        Min.   :0.0000   Length:7043
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.1621
##                                        3rd Qu.:0.0000
##                                        Max.   :1.0000
##
##   Dependents           tenure      PhoneService       MultipleLines
##  Length:7043        Min.   : 0.00   Length:7043        Length:7043
##  Class :character   1st Qu.: 9.00   Class :character   Class :character
##  Mode  :character   Median :29.00   Mode  :character   Mode  :character
##                     Mean   :32.37
##                     3rd Qu.:55.00
##                     Max.   :72.00
```

```
##
##  InternetService    OnlineSecurity     OnlineBackup       DeviceProtection
##  Length:7043        Length:7043        Length:7043        Length:7043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  TechSupport        StreamingTV        StreamingMovies      Contract
##  Length:7043        Length:7043        Length:7043        Length:7043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges
##  Length:7043        Length:7043        Min.   : 18.25   Min.   :  18.8
##  Class :character   Class :character   1st Qu.: 35.50   1st Qu.: 401.4
##  Mode  :character   Mode  :character   Median : 70.35   Median :1397.5
##                                        Mean   : 64.76   Mean   :2283.3
##                                        3rd Qu.: 89.85   3rd Qu.:3794.7
##                                        Max.   :118.75   Max.   :8684.8
##                                                         NA's   :11
##     Churn             Churn_int
##  Length:7043        Min.   :0.0000
##  Class :character   1st Qu.:0.0000
##  Mode  :character   Median :0.0000
##                     Mean   :0.2654
##                     3rd Qu.:1.0000
##                     Max.   :1.0000
##
```

```
#11 missing values in total charges
#%missing values :
11/7043
```

```
## [1] 0.001561834
```

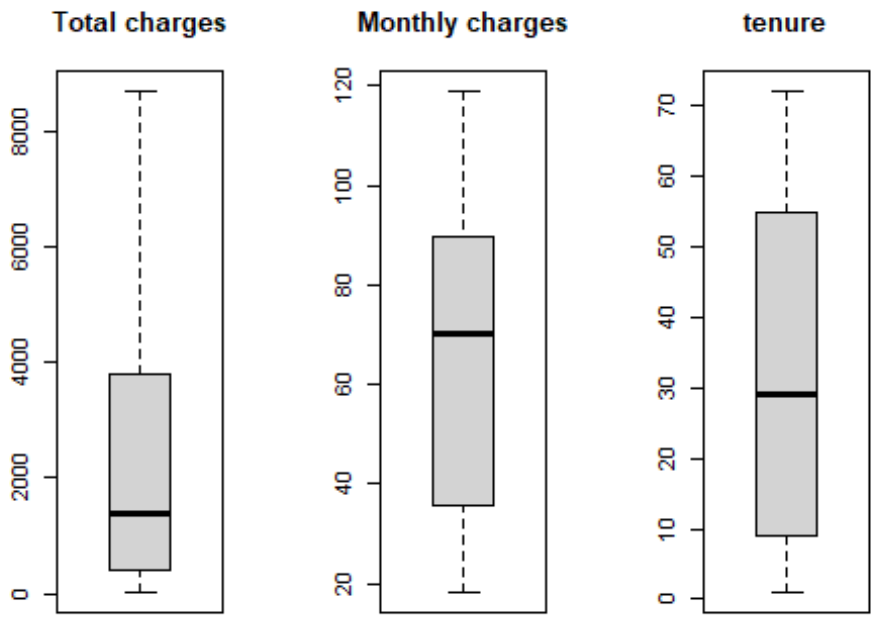
```
#which is very low.hence we discard the missing values.
```

```
data1=data[-which(is.na(data$TotalCharges)),]
View(data1)
summary(data1)
```

```
##   customerID          gender          SeniorCitizen      Partner
##  Length:7032        Length:7032        Min.   :0.0000   Length:7032
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.1624
```

```
##                                      3rd Qu.:0.0000
##                                      Max.   :1.0000
##   Dependents           tenure      PhoneService       MultipleLines
##  Length:7032        Min.   : 1.00   Length:7032        Length:7032
##  Class :character   1st Qu.: 9.00   Class :character   Class :character
##  Mode  :character   Median :29.00   Mode  :character   Mode  :character
##                     Mean   :32.42
##                     3rd Qu.:55.00
##                     Max.   :72.00
##  InternetService    OnlineSecurity     OnlineBackup       DeviceProtection
##  Length:7032        Length:7032        Length:7032        Length:7032
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  TechSupport        StreamingTV        StreamingMovies      Contract
##  Length:7032        Length:7032        Length:7032        Length:7032
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges
##  Length:7032        Length:7032        Min.   : 18.25   Min.   :  18.8
##  Class :character   Class :character   1st Qu.: 35.59   1st Qu.: 401.4
##  Mode  :character   Mode  :character   Median : 70.35   Median :1397.5
##                                        Mean   : 64.80   Mean   :2283.3
##                                        3rd Qu.: 89.86   3rd Qu.:3794.7
##                                        Max.   :118.75   Max.   :8684.8
##     Churn             Churn_int
##  Length:7032        Min.   :0.0000
##  Class :character   1st Qu.:0.0000
##  Mode  :character   Median :0.0000
##                     Mean   :0.2658
##                     3rd Qu.:1.0000
##                     Max.   :1.0000
```

```r
par(mfrow=c(1,3))
boxplot(data1$TotalCharges,main="Total charges")

boxplot(data1$MonthlyCharges,main="Monthly charges")

boxplot(data1$tenure,main="tenure")
```




```
#no outliers detected

par(mfrow=c(1,1))
#discarding columns leading to multicollinearity
data2=data1[,c(-1,-19,-8,-21,-10,-11,-12,-13,-14,-15)]

View(data2)
str(data2)

## 'data.frame':    7032 obs. of  12 variables:
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month"
"One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check"
"Bank transfer (automatic)" ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn_int       : int  0 0 1 0 1 1 0 0 1 0 ...

data2$gender=as.factor(data2$gender)
data2$SeniorCitizen=as.factor(data2$SeniorCitizen)
```

```
data2$Partner=as.factor(data2$Partner)
data2$Dependents=as.factor(data2$Dependents)
data2$PhoneService=as.factor(data2$PhoneService)
data2$InternetService =as.factor(data2$InternetService )
data2$Contract  =as.factor(data2$Contract)
data2$PaperlessBilling=as.factor(data2$PaperlessBilling)
data2$PaymentMethod  =as.factor(data2$PaymentMethod)
data2$Churn_int=as.factor(data2$Churn_int)
summary(data2)
```

```
##     gender     SeniorCitizen Partner    Dependents     tenure
PhoneService
##  Female:3483   0:5890        No :3639   No :4933   Min.   : 1.00   No :
680
##  Male  :3549   1:1142        Yes:3393   Yes:2099   1st Qu.: 9.00
Yes:6352
##                                                    Median :29.00
##                                                    Mean   :32.42
##                                                    3rd Qu.:55.00
##                                                    Max.   :72.00
##     InternetService           Contract    PaperlessBilling
##  DSL        :2416   Month-to-month:3875   No :2864
##  Fiber optic:3096   One year      :1472   Yes:4168
##  No         :1520   Two year      :1685
##
##
##
##                    PaymentMethod   TotalCharges    Churn_int
##  Bank transfer (automatic):1542   Min.   :  18.8   0:5163
##  Credit card (automatic)  :1521   1st Qu.: 401.4   1:1869
##  Electronic check         :2365   Median :1397.5
##  Mailed check             :1604   Mean   :2283.3
##                                   3rd Qu.:3794.7
##                                   Max.   :8684.8
```

```
(table(data2$Churn_int))
```

```
##
##    0    1
## 5163 1869
```

```
#imbalanced clases

library(ROSE)
```

```
## Warning: package 'ROSE' was built under R version 4.0.5
```

```
## Loaded ROSE 0.0-3
```

```
data3=ovun.sample(Churn_int~.,data = data2,method="over")$data

table(data3$Churn_int)
```

```
##
##    0    1
## 5163 5101
```

```
library(caret)
```

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

```
## Loading required package: ggplot2
```

```
index=createDataPartition(data3$Churn_int,p=0.7,list=FALSE)
tr=data3[index,] #Training set

te=data3[-index,] #Test Data
(table(tr$Churn_int))
```

```
##
##    0    1
## 3615 3571
```

```
model=glm(Churn_int~.,family = binomial,data=tr)
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.4
```

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

```
vif(model)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## gender           1.006342  1        1.003166
## SeniorCitizen    1.139156  1        1.067312
## Partner          1.381014  1        1.175166
## Dependents       1.295991  1        1.138416
## tenure          12.765614  1        3.572900
## PhoneService     1.396991  1        1.181944
## InternetService  2.495375  2        1.256852
## Contract         1.654696  2        1.134174
## PaperlessBilling 1.131299  1        1.063626
## PaymentMethod    1.404809  3        1.058286
## TotalCharges    14.695171  1        3.833428
```

```
#no multicolinearity
```

```
summary(model)
```

```
##
## Call:
## glm(formula = Churn_int ~ ., family = binomial, data = tr)
##
```

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.12908  -0.78990  -0.09104   0.76476   3.12101
##
## Coefficients:
##                                    Estimate Std. Error z value
Pr(>|z|)
## (Intercept)                         8.207e-01  1.450e-01   5.660 1.51e-
08 ***
## genderMale                          6.783e-02  5.868e-02   1.156
0.2477
## SeniorCitizen1                      1.489e-01  7.877e-02   1.890
0.0587 .
## PartnerYes                         -6.551e-02  6.959e-02  -0.941
0.3465
## DependentsYes                      -9.143e-02  7.879e-02  -1.160
0.2459
## tenure                             -5.458e-02  4.721e-03 -11.561  < 2e-
16 ***
## PhoneServiceYes                    -5.803e-01  1.175e-01  -4.941 7.79e-
07 ***
## InternetServiceFiber optic          9.297e-01  8.740e-02  10.637  < 2e-
16 ***
## InternetServiceNo                  -6.216e-01  1.109e-01  -5.606 2.07e-
08 ***
## ContractOne year                   -7.980e-01  8.931e-02  -8.935  < 2e-
16 ***
## ContractTwo year                   -1.591e+00  1.347e-01 -11.811  < 2e-
16 ***
## PaperlessBillingYes                 3.777e-01  6.586e-02   5.735 9.73e-
09 ***
## PaymentMethodCredit card (automatic) -2.150e-02  9.933e-02  -0.216
0.8287
## PaymentMethodElectronic check       4.255e-01  8.319e-02   5.114 3.15e-
07 ***
## PaymentMethodMailed check          -7.658e-02  1.005e-01  -0.762
0.4460
## TotalCharges                        3.085e-04  5.028e-05   6.136 8.48e-
10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9961.6  on 7185  degrees of freedom
## Residual deviance: 7073.3  on 7170  degrees of freedom
## AIC: 7105.3
##
## Number of Fisher Scoring iterations: 5
```

```r
#gender ,partner,dependents insignificant

library(lmtest)

## Warning: package 'lmtest' was built under R version 4.0.5

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

lrtest(model)

## Likelihood ratio test
##
## Model 1: Churn_int ~ gender + SeniorCitizen + Partner + Dependents +
tenure +
##     PhoneService + InternetService + Contract + PaperlessBilling +
##     PaymentMethod + TotalCharges
## Model 2: Churn_int ~ 1
##   #Df  LogLik  Df  Chisq Pr(>Chisq)
## 1  16 -3536.7
## 2   1 -4980.8 -15 2888.3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

model1=glm(Churn_int~gender+SeniorCitizen+Partner+Dependents+tenure+PhoneServ
ice+InternetService+Contract+PaperlessBilling+TotalCharges,data =tr,family =
binomial)

anova(model,model1,test="LRT") #Pvalue <0.05 include paymentmethod in the
model

## Analysis of Deviance Table
##
## Model 1: Churn_int ~ gender + SeniorCitizen + Partner + Dependents +
tenure +
##     PhoneService + InternetService + Contract + PaperlessBilling +
##     PaymentMethod + TotalCharges
## Model 2: Churn_int ~ gender + SeniorCitizen + Partner + Dependents +
tenure +
##     PhoneService + InternetService + Contract + PaperlessBilling +
##     TotalCharges
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      7170     7073.3
## 2      7173     7127.1 -3  -53.701  1.3e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2=glm(Churn_int~SeniorCitizen+tenure+PhoneService+InternetService+Contra
ct+PaperlessBilling+TotalCharges+PaymentMethod,data =tr,family = binomial)
lrtest(model2)
```

```
## Likelihood ratio test
##
## Model 1: Churn_int ~ SeniorCitizen + tenure + PhoneService +
InternetService +
##     Contract + PaperlessBilling + TotalCharges + PaymentMethod
## Model 2: Churn_int ~ 1
##   #Df  LogLik  Df  Chisq Pr(>Chisq)
## 1  13 -3539.2
## 2   1 -4980.8 -12 2883.3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tr$predprob=predict(model2,tr,type = "response")
#tr$predprob=round(fitted(model1),2)

library(InformationValue)
```

```
##
## Attaching package: 'InformationValue'
```

```
## The following objects are masked from 'package:caret':
##
##     confusionMatrix, precision, sensitivity, specificity
```

```
optimalCutoff(tr$Churn_int,tr$predprob) #0.45
```

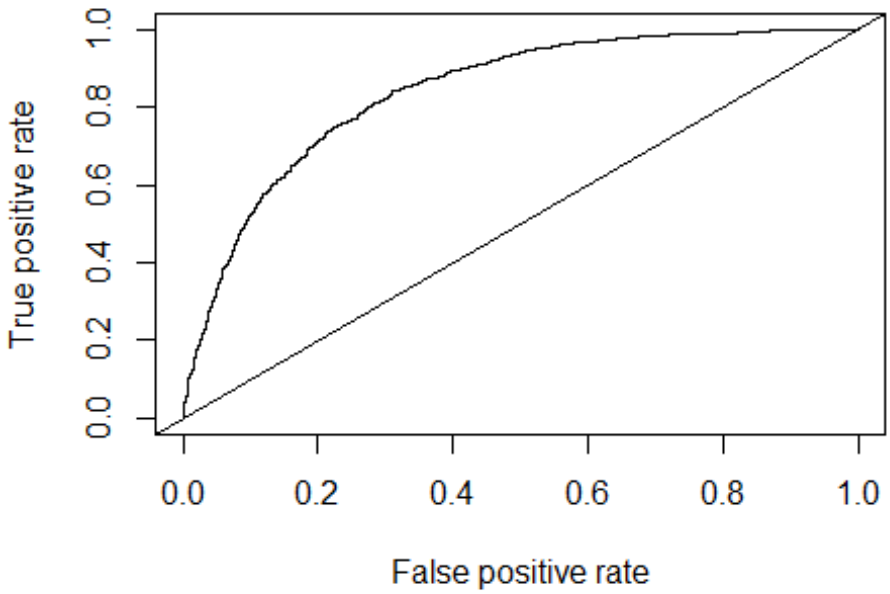
```
## [1] 0.4522829
```

```
tr$pred=ifelse(tr$predprob<0.45,0,1)

library(gmodels)
confusionMatrix(tr$Churn_int,tr$pred)
```

```
##      0    1
## 0 2436  547
## 1 1179 3091
```

```
accuracytr=(2436+3091)/(2436+3091+547+1179)
#accuracy=(2581+ 2931)/(2581+ 689+1034+ 2931)
sensitivitytr=(3091/(3091+547))
specificitytr=(2436/(2436+1179))
misclassification_ratetr=( 547+1179)/(2436+3091+547+1179)
#We are mis-identifying approx 23.79%

library(ROCR)
p=prediction(tr$predprob,tr$Churn_int)
```

```
perf=performance(p,"tpr","fpr")
plot(perf); abline(0,1)
```




```
auc=performance(p,"auc")
auc@y.values #84.16%

## [[1]]
## [1] 0.841624

#Checking on test data
te$predprob=predict(model2,te,type="response")
te$pred=ifelse(te$predprob<0.45,0,1)
confusionMatrix(te$Churn_int,te$pred)

##      0    1
## 0 1079  261
## 1  469 1290

accuracy_te=(1079+1290)/(469+261+1079+1290)
sensitivity_te=(1290)/(1290+261 )
specificity_te=1079/(1079+469)
misclassification_rate_te=(261+469)/(469+261+1079+1290)
```