CA2-NeerajWamanChavan

Neeraj Waman Chavan

24/12/2019

Introduction

setwd("D:/Dublin Business School/Semester 1/Statistics and Mathematics for Data Analytics/Assignments/CA2")  
#Loading the dataset  
data<-read.csv("Bank\_Data.csv")  
head(data)

## RowNumber CustomerId Surname CreditScore Geography Gender Age Tenure  
## 1 1 15634602 Hargrave 619 France Female 42 2  
## 2 2 15647311 Hill 608 Spain Female 41 1  
## 3 3 15619304 Onio 502 France Female 42 8  
## 4 4 15701354 Boni 699 France Female 39 1  
## 5 5 15737888 Mitchell 850 Spain Female 43 2  
## 6 6 15574012 Chu 645 Spain Male 44 8  
## Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited  
## 1 0.00 1 1 1 101348.88 1  
## 2 83807.86 1 0 1 112542.58 0  
## 3 159660.80 3 1 0 113931.57 1  
## 4 0.00 2 0 0 93826.63 0  
## 5 125510.82 1 1 1 79084.10 0  
## 6 113755.78 2 1 0 149756.71 1

Defining output and input variables

y <- data$Exited #dependent variable  
x1 <- data$CreditScore #independent variable  
x2 <- data$Geography #independent variable  
x3 <- data$Gender #independent variable  
x4 <- data$Age #independent variable  
x5 <- data$Tenure #independent variable  
x6 <- data$Balance #independent variable  
x7 <- data$NumOfProducts #independent variable  
x8 <- data$HasCrCard #independent variable  
x9 <- data$IsActiveMember #independent variable  
x10 <- data$EstimatedSalary #independent variable

Cleaning the dataset and loading it into a data frame

df=data.frame( x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,y)  
dataset=na.omit(df)

Splitting the dataset into 80% trainset and 20% testset

set.seed(1390) #for verifying answers  
n=nrow(dataset) #storing number of rows of dataset in a variable  
indexes = sample(n,n\*(80/100)) #splitting it 80-20  
trainset = dataset[indexes,] #defining the training set  
testset = dataset[-indexes,] #defining the test set

Fitting the model using trainset

trainset.glm <- glm(trainset$y ~.,trainset, family="binomial") #using glm to fit the model  
summary(trainset.glm) #displaying the summary of the fitted model to check what variables are affecting the output

##   
## Call:  
## glm(formula = trainset$y ~ ., family = "binomial", data = trainset)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2941 -0.6632 -0.4584 -0.2699 2.9051   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.341e+00 2.718e-01 -12.292 < 2e-16 \*\*\*  
## x1 -7.180e-04 3.122e-04 -2.300 0.0214 \*   
## x2Germany 7.869e-01 7.568e-02 10.397 < 2e-16 \*\*\*  
## x2Spain 5.689e-02 7.804e-02 0.729 0.4660   
## x3Male -5.170e-01 6.059e-02 -8.532 < 2e-16 \*\*\*  
## x4 7.200e-02 2.862e-03 25.162 < 2e-16 \*\*\*  
## x5 -1.392e-02 1.045e-02 -1.332 0.1828   
## x6 2.234e-06 5.682e-07 3.932 8.43e-05 \*\*\*  
## x7 -9.091e-02 5.198e-02 -1.749 0.0803 .   
## x8 -2.980e-02 6.629e-02 -0.450 0.6531   
## x9 -1.089e+00 6.417e-02 -16.979 < 2e-16 \*\*\*  
## x10 6.740e-07 5.253e-07 1.283 0.1995   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8140.4 on 7999 degrees of freedom  
## Residual deviance: 6909.1 on 7988 degrees of freedom  
## AIC: 6933.1  
##   
## Number of Fisher Scoring iterations: 5

Prediction

phat=predict(trainset.glm , testset, type='response')  
pred=rep(0, length(phat))  
pred[phat>=0.5]=1  
pred

## [1] 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0  
## [35] 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0  
## [69] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0  
## [103] 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [137] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0  
## [171] 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [205] 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0  
## [239] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [273] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0  
## [307] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [341] 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [375] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1  
## [409] 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1  
## [443] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0  
## [477] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0  
## [511] 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [545] 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0  
## [579] 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0  
## [613] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [647] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [681] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0  
## [715] 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0  
## [749] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0  
## [783] 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [817] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0  
## [851] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0  
## [885] 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0  
## [919] 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [953] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0  
## [987] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1021] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1055] 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0  
## [1089] 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1123] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0  
## [1157] 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0  
## [1191] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1225] 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1259] 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1293] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0  
## [1327] 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0  
## [1361] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0  
## [1395] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 1 0 0  
## [1429] 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1  
## [1463] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1497] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0  
## [1531] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0  
## [1565] 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0  
## [1599] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1  
## [1633] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0  
## [1667] 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1701] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0  
## [1735] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0  
## [1769] 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1  
## [1803] 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1837] 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0  
## [1871] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1905] 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0  
## [1939] 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [1973] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0

Confusion Matrix to calculate the accuracy of the model

actual=testset$y  
conf\_mat=table(pred,actual)  
accuracy=mean(pred==actual)  
conf\_mat

## actual  
## pred 0 1  
## 0 1550 299  
## 1 62 89

accuracy

## [1] 0.8195