DeSimone\_MS64060\_Assignment 3

Heather DeSimone

3/5/2022

##First I have loaded in my data frame and called a summary of the information.

DF=read.csv("C:/Users/hdesi/Desktop/MBA/Machine Learning/UniversalBank2.csv")  
library(dplyr)

##   
## 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(caret)

## Warning: package 'caret' was built under R version 4.1.2

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.1.2

## Loading required package: lattice

library(class)  
library(ISLR)

## Warning: package 'ISLR' was built under R version 4.1.1

DF <- DF %>% relocate(Personal.Loan, .after = CreditCard)  
summary(DF)

## ID Age Experience Income ZIP.Code   
## Min. : 1 Min. :23.00 Min. :-3.0 Min. : 8.00 Min. : 9307   
## 1st Qu.:1251 1st Qu.:35.00 1st Qu.:10.0 1st Qu.: 39.00 1st Qu.:91911   
## Median :2500 Median :45.00 Median :20.0 Median : 64.00 Median :93437   
## Mean :2500 Mean :45.34 Mean :20.1 Mean : 73.77 Mean :93153   
## 3rd Qu.:3750 3rd Qu.:55.00 3rd Qu.:30.0 3rd Qu.: 98.00 3rd Qu.:94608   
## Max. :5000 Max. :67.00 Max. :43.0 Max. :224.00 Max. :96651   
## Family CCAvg Education Mortgage   
## Min. :1.000 Min. : 0.000 Min. :1.000 Min. : 0.0   
## 1st Qu.:1.000 1st Qu.: 0.700 1st Qu.:1.000 1st Qu.: 0.0   
## Median :2.000 Median : 1.500 Median :2.000 Median : 0.0   
## Mean :2.396 Mean : 1.938 Mean :1.881 Mean : 56.5   
## 3rd Qu.:3.000 3rd Qu.: 2.500 3rd Qu.:3.000 3rd Qu.:101.0   
## Max. :4.000 Max. :10.000 Max. :3.000 Max. :635.0   
## Securities.Account CD.Account Online CreditCard   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000   
## Median :0.0000 Median :0.0000 Median :1.0000 Median :0.000   
## Mean :0.1044 Mean :0.0604 Mean :0.5968 Mean :0.294   
## 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.000   
## Personal.Loan   
## Min. :0.000   
## 1st Qu.:0.000   
## Median :0.000   
## Mean :0.096   
## 3rd Qu.:0.000   
## Max. :1.000

##I have converted a few attributes over to factors - these attributes classify a yes (1) or no (0) response.I have called a summary to check my work.

DF$Personal.Loan=as.factor(DF$Personal.Loan)  
DF$Securities.Account=as.factor(DF$Securities.Account)  
DF$CD.Account=as.factor(DF$CD.Account)  
DF$Online=as.factor(DF$Online)  
DF$CreditCard=as.factor(DF$CreditCard)  
summary(DF)

## ID Age Experience Income ZIP.Code   
## Min. : 1 Min. :23.00 Min. :-3.0 Min. : 8.00 Min. : 9307   
## 1st Qu.:1251 1st Qu.:35.00 1st Qu.:10.0 1st Qu.: 39.00 1st Qu.:91911   
## Median :2500 Median :45.00 Median :20.0 Median : 64.00 Median :93437   
## Mean :2500 Mean :45.34 Mean :20.1 Mean : 73.77 Mean :93153   
## 3rd Qu.:3750 3rd Qu.:55.00 3rd Qu.:30.0 3rd Qu.: 98.00 3rd Qu.:94608   
## Max. :5000 Max. :67.00 Max. :43.0 Max. :224.00 Max. :96651   
## Family CCAvg Education Mortgage   
## Min. :1.000 Min. : 0.000 Min. :1.000 Min. : 0.0   
## 1st Qu.:1.000 1st Qu.: 0.700 1st Qu.:1.000 1st Qu.: 0.0   
## Median :2.000 Median : 1.500 Median :2.000 Median : 0.0   
## Mean :2.396 Mean : 1.938 Mean :1.881 Mean : 56.5   
## 3rd Qu.:3.000 3rd Qu.: 2.500 3rd Qu.:3.000 3rd Qu.:101.0   
## Max. :4.000 Max. :10.000 Max. :3.000 Max. :635.0   
## Securities.Account CD.Account Online CreditCard Personal.Loan  
## 0:4478 0:4698 0:2016 0:3530 0:4520   
## 1: 522 1: 302 1:2984 1:1470 1: 480   
##   
##   
##   
##

##Question A ##I will now separate my data into training and validating sets - training = 60% and validation = 40%. ##I have also created my pivot table.

Train\_Index = createDataPartition(DF$Personal.Loan,p=0.6, list=FALSE)  
Train.df=DF[Train\_Index,]  
Validation.df=DF[-Train\_Index,]   
  
mytable <- xtabs(~ CreditCard+Online+Personal.Loan, data=Train.df)  
ftable(mytable)

## Personal.Loan 0 1  
## CreditCard Online   
## 0 0 766 79  
## 1 1141 122  
## 1 0 321 34  
## 1 484 53

##Question B ##The probability that a customer will accept a loan offer based condionally that they have a credit card and online account is roughly 10% (.0996)

##Question C ##Creating my 2 new pivot tables

table(Personal.Loan=Train.df$Personal.Loan, Online=Train.df$Online)

## Online  
## Personal.Loan 0 1  
## 0 1087 1625  
## 1 113 175

table(Personal.Loan=Train.df$Personal.Loan, CreditCard=Train.df$CreditCard)

## CreditCard  
## Personal.Loan 0 1  
## 0 1907 805  
## 1 201 87

##Question D

##i. P(CC = 1 | Loan = 1) (the proportion of credit card holders among the loan acceptors) ## Answer is 92/(196+92) = .319 = 32%

##ii. P(Online = 1 | Loan = 1) (the proportion of Online users among the loan acceptors) ## Answer is 172/(116+172) = .597 = 60%

##iii. P(Loan = 1) (the proportion of loan acceptors)  
## Answer is (196+92)/(1917+795+196+92) = 288/3000 = .096 = 10%

##iv. P(CC = 1 | Loan = 0)  
##Answer is 795/(795+1917) = .293 = 29%

##v. P(Online = 1 | Loan = 0) ## Answer is 1594/(1118+1594) = .587 = 59%

##vi. P(Loan = 0) ## Answer is (1917+795)/(1917+795+196+92) = .904 = 90%

##Question E ##P(Loan = 1 | CC = 1, Online = 1). ##P(Loan = 1) = .319\*.597 = .19

##Question F ##The pivot table is more accurate because there are more variables used in the prediction. The Naive Bayes assumes that each prediction is independent from each variable.

##Question G ##Running Naive Bayes on the data

library(e1071)

## Warning: package 'e1071' was built under R version 4.1.2

nb.model<-naiveBayes (Personal.Loan~CreditCard+Online, data=Train.df)  
To\_Predict=data.frame(CreditCard='1', Online='1')  
predict(nb.model,To\_Predict,type='raw')

## 0 1  
## [1,] 0.9012268 0.09877325

##The above running of the naive bayes on my data is very close to the prediction I made in Question B. Question E has a very different answer than B and G. I would conclude that 10% is the correct prediction.