Assignment\_3

Nikhila Reddy

2024-03-10

#Importing the dataset

Universal\_Bank\_1 <- read.csv("C:\\Users\\Nikhi\\Downloads\\Universal\_Bank\_1.csv")

#Loading the required packages.

library("caret")

## Loading required package: ggplot2

## Loading required package: lattice

##Loading required package: ggplot2 ##Loading required package: lattice

library("ISLR")  
library("ggplot2")  
library("class")  
library("lattice")  
library("reshape2")

#install.packages(“melt”)

library("melt")

## Warning: package 'melt' was built under R version 4.3.3

#Transforming to factor variable.

Universal\_Bank\_1$Personal.Loan <- as.factor(Universal\_Bank\_1$Personal.Loan)  
Universal\_Bank\_1$Online <- as.factor(Universal\_Bank\_1$Online)  
Universal\_Bank\_1$CreditCard <- as.factor(Universal\_Bank\_1$CreditCard)

#Checking the summary of the dataset.

summary(Universal\_Bank\_1)

## 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 Personal.Loan  
## Min. :1.000 Min. : 0.000 Min. :1.000 Min. : 0.0 0:4520   
## 1st Qu.:1.000 1st Qu.: 0.700 1st Qu.:1.000 1st Qu.: 0.0 1: 480   
## 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 0:2016 0:3530   
## 1st Qu.:0.0000 1st Qu.:0.0000 1:2984 1:1470   
## Median :0.0000 Median :0.0000   
## Mean :0.1044 Mean :0.0604   
## 3rd Qu.:0.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000

#The above data represents summary for the given dataset.

#Separating the entire dataset into training and testing subsets, #Allocating 60% designated for training and 40% for validation.

set.seed(23)

Split\_Index <- createDataPartition(Universal\_Bank\_1$Personal.Loan, p=0.6, list = FALSE)  
Train\_data <- Universal\_Bank\_1[Split\_Index,]  
Validation\_data <- Universal\_Bank\_1[-Split\_Index,]

#The above data is now splitted into training (60%) and testing (40%) sets.

dim(Train\_data)

## [1] 3000 14

dim(Validation\_data)

## [1] 2000 14

normalising1 <- preProcess(Train\_data[, -c(10,13:14)],method = c("center","scale"))  
Train\_data1 <- predict(normalising1,Train\_data)  
Validation\_data1 <- predict(normalising1,Validation\_data)

A)Create a pivot table for the training data with Online as a column variable, and Loan as a secondary row variable. The values inside the table should convey tye count. In R use functions melt() ad cast(), or function table(). In Python, use panda dataframe methods melt() and pivot()

PTable\_1 <- ftable(Train\_data1[,c(14,10,13)])  
PTable\_1

## Online 0 1  
## CreditCard Personal.Loan   
## 0 0 773 1127  
## 1 82 114  
## 1 0 315 497  
## 1 39 53

#Here 14- CreditCard, 10- Personal.Loan, 13- Online

1. Consider the task of classifying a customer who owns a bank credit card and is actively using online banking services. Looking at the pivot table, What is the probability that this customer will accept the loan offer?

Ans) Utilizing the data from the pivot table, we can calculate the likelihood of the customer acceting the loan offer as 52/(52+503), resulting in a probability of 0.096.

1. Create two seperate pivot tables for the training data. One will have Loan (rows) as a function of Online(Columns) and the other will have Loan(rows) as a function of CC.

melt\_1 <- melt(Train\_data1,id=c("Personal.Loan"),variable="Online")

## Warning: attributes are not identical across measure variables; they will be  
## dropped

melt\_2 <- melt(Train\_data1,id=c("Personal.Loan"),variable="CreditCard")

## Warning: attributes are not identical across measure variables; they will be  
## dropped

cast\_1 = dcast(melt\_1, Personal.Loan~Online)

## Aggregation function missing: defaulting to length

cast\_2 <- dcast(melt\_2, Personal.Loan~CreditCard)

## Aggregation function missing: defaulting to length

1. Compute the following quantities [P(A | B) means “the probability ofA given B”]:
2. P(CC = 1 | Loan = 1) (the proportion of credit card holders among the loan acceptors)
3. P(Online = 1 | Loan = 1)
4. P(Loan = 1) (the proportion of loan acceptors)
5. P(CC = 1 | Loan = 0)
6. P(Online = 1 | Loan = 0)
7. P(Loan = 0)

ftable(Train\_data1[,c(10,13)])

## Online 0 1  
## Personal.Loan   
## 0 1088 1624  
## 1 121 167

ftable(Train\_data1[,c(10,14)])

## CreditCard 0 1  
## Personal.Loan   
## 0 1900 812  
## 1 196 92

1.P(CC = 1 | Loan = 1) = (92/92+196) = 0.319 2.P(Online = 1 | Loan = 1) = (167/167+121) = 0.579 3.P(Loan = 1) = (288/288+2712) = 0.096 4.P(CC = 1 | Loan = 0) = (812/812+1900) = 0.299 5.P(Online = 1 | Loan = 0) = (1624/ 1624+1088) = 0.598 6.P(Loan = 0) = (2712/ 2712+288) = 0.904

E). Use the quantities computed above to compute the naive Bayes probability P(Loan = 1 | CC = 1, Online = 1) Ans) (0.319*0.579*0.096)/(0.319*0.579*0.096) + (0.299*0.598*0.904)= 0.098

F).Compare this value with the one obtained from the pivot table in (B). Which is a more accurate estimate? Ans) In section B, we derived a probability value of 0.096, and in the previous question, we computed a probability value of 0.098. Despite these slight differences, it’s essential to highlight that in part B, we incorporated a broader range of dependent information. Consequently, we can confidently state that the value obtained in part B offers a more precise and nuanced representation of the dataset

G). Which of the entries in this table are needed for computing P(Loan = 1 | CC = 1, Online = 1)? Run naive Bayes on the data. Examine the model output on training data, and find the entry that corresponds to P(Loan = 1 | CC = 1, Online = 1). Compare this to the number you obtained in (E)

#install.packages(“naivebayes”)

library("naivebayes")

## Warning: package 'naivebayes' was built under R version 4.3.3

## naivebayes 0.9.7 loaded

naive\_b <- naive\_bayes(Personal.Loan~Online+CreditCard,data=Train\_data1)  
naive\_b

##   
## ================================== Naive Bayes ==================================   
##   
## Call:   
## naive\_bayes.formula(formula = Personal.Loan ~ Online + CreditCard,   
## data = Train\_data1)  
##   
## ---------------------------------------------------------------------------------   
##   
## Laplace smoothing: 0  
##   
## ---------------------------------------------------------------------------------   
##   
## A priori probabilities:   
##   
## 0 1   
## 0.904 0.096   
##   
## ---------------------------------------------------------------------------------   
##   
## Tables:   
##   
## ---------------------------------------------------------------------------------   
## ::: Online (Bernoulli)   
## ---------------------------------------------------------------------------------   
##   
## Online 0 1  
## 0 0.4011799 0.4201389  
## 1 0.5988201 0.5798611  
##   
## ---------------------------------------------------------------------------------   
## ::: CreditCard (Bernoulli)   
## ---------------------------------------------------------------------------------   
##   
## CreditCard 0 1  
## 0 0.7005900 0.6805556  
## 1 0.2994100 0.3194444  
##   
## ---------------------------------------------------------------------------------

-The probability of a customer accepting the loan, utilizing a credit card, and participating in online banking is 0.096 as per the Naive Bayes Model. This outcome closely aligns with the value derived in section E of our analysis.