Assignment\_3(FML)

“ADITI”

2024-03-10

# read the dataset  
bank\_data <- read.csv("UniversalBank.csv")  
bank\_data

## ID Age Experience Income ZIP.Code Family CCAvg Education Mortgage  
## 1 1 25 1 49 91107 4 1.60 1 0  
## 2 2 45 19 34 90089 3 1.50 1 0  
## 3 3 39 15 11 94720 1 1.00 1 0  
## 4 4 35 9 100 94112 1 2.70 2 0  
## 5 5 35 8 45 91330 4 1.00 2 0  
## 6 6 37 13 29 92121 4 0.40 2 155  
## 7 7 53 27 72 91711 2 1.50 2 0  
## 8 8 50 24 22 93943 1 0.30 3 0  
## 9 9 35 10 81 90089 3 0.60 2 104  
## 10 10 34 9 180 93023 1 8.90 3 0  
## 11 11 65 39 105 94710 4 2.40 3 0  
4 1.70 2 0  
## 388 388 31 5 82 95482 4 1.80 2 0  
## 389 389 54 30 100 95814 4 3.40 3 0  
## 390 390 45 20 155 90024 1 7.00 1 0  
## 391 391 45 19 45 92521 1 0.20 1 0  
## 392 392 58 32 9 94080 3 0.30 3 0  
## 393 393 54 29 48 91709 4 1.80 3 0  
## 394 394 53 28 18 90095 4 0.10 3 109  
## 395 395 33 9 80 91311 4 3.40 1 0  
## 396 396 60 35 64 94509 2 2.80 1 0  
## 397 397 50 24 29 93023 4 0.10 1 0  
## 398 398 26 2 48 90503 3 0.70 2 0  
## 399 399 54 30 23 94608 2 0.40 1 0  
## 400 400 28 3 84 90024 4 0.20 1 0  
## 401 401 36 10 179 94542 3 6.60 1 0  
## 402 402 29 2 30 95747 4 1.50 2 112  
## 403 403 54 28 93 91604 1 4.90 1 133  
## 404 404 55 30 39 92647 2 1.90 2 0  
## 405 405 61 36 60 92866 3 0.50 2 182  
## 406 406 36 11 133 90245 1 3.80 1 290  
## 407 407 45 19 125 92354 1 2.40 1 0  
## 408 408 64 40 58 93437 1 1.80 3 0  
## 409 409 60 36 89 91745 2 2.80 1 0  
## 410 410 49 22 82 90019 1 2.67 2 125  
## 411 411 47 23 110 94111 2 3.30 1 0  
## 412 412 60 36 54 92182 4 2.30 3 0

head(bank\_data)#output gives first 5 rows of the dataset

## ID Age Experience Income ZIP.Code Family CCAvg Education Mortgage  
## 1 1 25 1 49 91107 4 1.6 1 0  
## 2 2 45 19 34 90089 3 1.5 1 0  
## 3 3 39 15 11 94720 1 1.0 1 0  
## 4 4 35 9 100 94112 1 2.7 2 0  
## 5 5 35 8 45 91330 4 1.0 2 0  
## 6 6 37 13 29 92121 4 0.4 2 155  
## Personal.Loan Securities.Account CD.Account Online CreditCard  
## 1 0 1 0 0 0  
## 2 0 1 0 0 0  
## 3 0 0 0 0 0  
## 4 0 0 0 0 0  
## 5 0 0 0 0 1  
## 6 0 0 0 1 0

tail(bank\_data)#output gives last 5 rows of the dataset

## ID Age Experience Income ZIP.Code Family CCAvg Education Mortgage  
## 4995 4995 64 40 75 94588 3 2.0 3 0  
## 4996 4996 29 3 40 92697 1 1.9 3 0  
## 4997 4997 30 4 15 92037 4 0.4 1 85  
## 4998 4998 63 39 24 93023 2 0.3 3 0  
## 4999 4999 65 40 49 90034 3 0.5 2 0  
## 5000 5000 28 4 83 92612 3 0.8 1 0  
## Personal.Loan Securities.Account CD.Account Online CreditCard  
## 4995 0 0 0 1 0  
## 4996 0 0 0 1 0  
## 4997 0 0 0 1 0  
## 4998 0 0 0 0 0  
## 4999 0 0 0 1 0  
## 5000 0 0 0 1 1

nrow(bank\_data)#there are 5000 rows

## [1] 5000

library(e1071)  
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(ISLR)

Partition the data into training (60%) and validation (40%) sets.

bank\_data$Personal.Loan = as.factor(bank\_data$Personal.Loan)  
bank\_data$Online = as.factor(bank\_data$Online)  
bank\_data$CreditCard = as.factor(bank\_data$CreditCard)  
set.seed(111)  
train\_index <- sample(row.names(bank\_data), 0.6\*dim(bank\_data)[1])   
valid\_index <- setdiff(row.names(bank\_data), train\_index)   
Train\_data <- bank\_data[train\_index, ]  
valid\_data <- bank\_data[valid\_index, ]

A. Create a pivot table for the training data with Online as a column variable, CC as a row variable, and Loan as a secondary row variable. The values inside the table should convey the count. In R use functions melt() and cast(),

library(reshape2)  
melt.data <- melt(Train\_data,id.vars = c("CreditCard","Personal.Loan"),variable = "Online")

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

pivot\_tb <- dcast(melt.data, CreditCard + Personal.Loan ~ Online)

## Aggregation function missing: defaulting to length

pivot\_tb

## CreditCard Personal.Loan ID Age Experience Income ZIP.Code Family CCAvg  
## 1 0 0 1896 1896 1896 1896 1896 1896 1896  
## 2 0 1 205 205 205 205 205 205 205  
## 3 1 0 812 812 812 812 812 812 812  
## 4 1 1 87 87 87 87 87 87 87  
## Education Mortgage Securities.Account CD.Account Online  
## 1 1896 1896 1896 1896 1896  
## 2 205 205 205 205 205  
## 3 812 812 812 812 812  
## 4 87 87 87 87 87

pivot\_tb[,c(1:2,14)]

## CreditCard Personal.Loan Online  
## 1 0 0 1896  
## 2 0 1 205  
## 3 1 0 812  
## 4 1 1 87

B.Consider the task of classifying a customer who owns a bank credit card and is actively usingonline banking services. Looking at the pivot table, what is the probability that this customerwill accept the loan offer? [This is the probability of loan acceptance (Loan = 1) conditional onhaving a bank credit card (CC = 1) and being an active user of online banking services (Online= 1)] sOLUTION: The only record where the credit card value is 1 , loan acceptance is 1 as well as we have a value for online therefore if we add all the online column values and divide by the value where all values are 1 or the assign value we can get the probability

94/3017

## [1] 0.03115678

C.Create two separate 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.

melted.data1 <- melt(Train\_data,id.vars = c("Personal.Loan"), var = "Online")

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

piv\_1 <- dcast(melted.data1, Personal.Loan ~ Online)

## Aggregation function missing: defaulting to length

piv\_1

## Personal.Loan ID Age Experience Income ZIP.Code Family CCAvg Education  
## 1 0 2708 2708 2708 2708 2708 2708 2708 2708  
## 2 1 292 292 292 292 292 292 292 292  
## Mortgage Securities.Account CD.Account Online CreditCard  
## 1 2708 2708 2708 2708 2708  
## 2 292 292 292 292 292

piv\_1[,c(1,13)]

## Personal.Loan Online  
## 1 0 2708  
## 2 1 292

melted.data2 <- melt(Train\_data,id.vars = "Personal.Loan", var = "CreditCard")

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

piv\_2 <- dcast(melted.data2, Personal.Loan ~ CreditCard)

## Aggregation function missing: defaulting to length

piv\_2

## Personal.Loan ID Age Experience Income ZIP.Code Family CCAvg Education  
## 1 0 2708 2708 2708 2708 2708 2708 2708 2708  
## 2 1 292 292 292 292 292 292 292 292  
## Mortgage Securities.Account CD.Account Online CreditCard  
## 1 2708 2708 2708 2708 2708  
## 2 292 292 292 292 292

piv\_2[,c(1,14)]

## Personal.Loan CreditCard  
## 1 0 2708  
## 2 1 292

D.Compute the following quantities [P(A | B) means “the probability ofA given B”]: i. P(CC = 1 | Loan = 1) (the proportion of credit card holders among the loan acceptors) ii. P(Online = 1 | Loan = 1) iii. P(Loan = 1) (the proportion of loan acceptors) iv. P(CC = 1 | Loan = 0) v. P(Online = 1 | Loan = 0) vi. P(Loan = 0).

t1 <- table(Train\_data[,c(14,10)])  
t1

## Personal.Loan  
## CreditCard 0 1  
## 0 1896 205  
## 1 812 87

t2 <- table(Train\_data[,c(13,10)])  
t2

## Personal.Loan  
## Online 0 1  
## 0 1112 112  
## 1 1596 180

t3 <- table(Train\_data[,10])  
t3

##   
## 0 1   
## 2708 292

1. P(CC = 1 | Loan = 1) (the proportion of credit card holders among the loan acceptors)

# from t1 we can calculate   
94/(94 + 209)

## [1] 0.310231

1. P(Online = 1 | Loan = 1)

#from t2 we calculate  
181/(181 + 122)

## [1] 0.5973597

1. P(Loan = 1) (the proportion of loan acceptors)

#from t3 we calculate  
276/(276 + 2755)

## [1] 0.09105906

iv.P(CC = 1 | Loan = 0)

816/(816 + 1939)

## [1] 0.2961887

1. P(Online = 1 | Loan = 0)

1641/(1641 + 1114)

## [1] 0.5956443

1. P(Loan = 0)

2755/(2755 + 276)

## [1] 0.9089409

E. Use the quantities computed above to compute the naive Bayes probability P(Loan = 1 | CC = 1, Online = 1).

((94/(94 + 209))\*(181/(181 + 122))\*(276/(276 + 2755)))/(((94/(94 + 209))\*(181/(181 + 122))\*(276/(276 + 2755)))+((816/(816 + 1939))\*(1641/(1641 + 1114))\*2755/(2755 + 276)))

## [1] 0.09521364

F. Compare this value with the one obtained from the pivot table in (B). Which is a more accurate estimate?

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).

nb\_train = Train\_data[,c(10,13:14)]  
nb\_valid = valid\_data[,c(10,13:14)]  
nb\_model = naiveBayes(Personal.Loan~.,data=nb\_train)  
print(nb\_model)

##   
## Naive Bayes Classifier for Discrete Predictors  
##   
## Call:  
## naiveBayes.default(x = X, y = Y, laplace = laplace)  
##   
## A-priori probabilities:  
## Y  
## 0 1   
## 0.90266667 0.09733333   
##   
## Conditional probabilities:  
## Online  
## Y 0 1  
## 0 0.4106352 0.5893648  
## 1 0.3835616 0.6164384  
##   
## CreditCard  
## Y 0 1  
## 0 0.7001477 0.2998523  
## 1 0.7020548 0.2979452

(0.29)\*(0.61)\*(0.09)/ (0.29\*0.61\*0.09+0.29\*0.59\*0.90)

## [1] 0.093702