Assignment\_3

Hemani Panchmatiya

2025-10-03

#Introduction: This analysis is based on Universal Bank data set to predict loan acceptance of customers based on two factors : Credit card holding and Online banking user. Pivot tables are used to examine the frequencies, relationship of variables. Naive Bayes classifier helps in estimating the probabilities under assumptions of conditional independence. Comparing both methods help in understanding the accuracy of model and patterns in customer behavior.

#Loading the packages

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(ISLR)  
library(e1071)

##   
## Attaching package: 'e1071'

## The following object is masked from 'package:ggplot2':  
##   
## element

library(ggplot2)  
library(readr)  
library(naivebayes)

## naivebayes 1.0.0 loaded

## For more information please visit:

## https://majkamichal.github.io/naivebayes/

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(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

#Importing the dataset UniversalBank

Mydata <- read\_csv("C:/Users/panch/OneDrive/Desktop/FML Assignments/Assignmnet 2/UniversalBank.csv")

## Rows: 5000 Columns: 14  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (14): ID, Age, Experience, Income, ZIP Code, Family, CCAvg, Education, M...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#Explore the dataset UniversalBank

str(Mydata)

## spc\_tbl\_ [5,000 × 14] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ID : num [1:5000] 1 2 3 4 5 6 7 8 9 10 ...  
## $ Age : num [1:5000] 25 45 39 35 35 37 53 50 35 34 ...  
## $ Experience : num [1:5000] 1 19 15 9 8 13 27 24 10 9 ...  
## $ Income : num [1:5000] 49 34 11 100 45 29 72 22 81 180 ...  
## $ ZIP Code : num [1:5000] 91107 90089 94720 94112 91330 ...  
## $ Family : num [1:5000] 4 3 1 1 4 4 2 1 3 1 ...  
## $ CCAvg : num [1:5000] 1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...  
## $ Education : num [1:5000] 1 1 1 2 2 2 2 3 2 3 ...  
## $ Mortgage : num [1:5000] 0 0 0 0 0 155 0 0 104 0 ...  
## $ Personal Loan : num [1:5000] 0 0 0 0 0 0 0 0 0 1 ...  
## $ Securities Account: num [1:5000] 1 1 0 0 0 0 0 0 0 0 ...  
## $ CD Account : num [1:5000] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Online : num [1:5000] 0 0 0 0 0 1 1 0 1 0 ...  
## $ CreditCard : num [1:5000] 0 0 0 0 1 0 0 1 0 0 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ID = col\_double(),  
## .. Age = col\_double(),  
## .. Experience = col\_double(),  
## .. Income = col\_double(),  
## .. `ZIP Code` = col\_double(),  
## .. Family = col\_double(),  
## .. CCAvg = col\_double(),  
## .. Education = col\_double(),  
## .. Mortgage = col\_double(),  
## .. `Personal Loan` = col\_double(),  
## .. `Securities Account` = col\_double(),  
## .. `CD Account` = col\_double(),  
## .. Online = col\_double(),  
## .. CreditCard = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

#Checking if the data has any missing values

sum(is.na(Mydata))

## [1] 0

#Splitting the data into Training (60%) and validation(40%)

set.seed(123)  
train\_index <- createDataPartition(Mydata$`Personal Loan`, p = 0.6, list = FALSE)  
train\_data <- Mydata[train\_index, ]  
valid\_data <- Mydata[-train\_index, ]  
nrow(train\_data)

## [1] 3000

nrow(valid\_data)

## [1] 2000

#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 using table function in R.

#Creating Pivot table using function table()

Pivottable <- ftable(train\_data$CreditCard, train\_data$`Personal Loan`, train\_data$Online)  
Pivottable

## 0 1  
##   
## 0 0 785 1145  
## 1 65 122  
## 1 0 317 475  
## 1 34 57

#Insights: A pivot table is created using ftable(). The rows represents whether the customers has a credit card, secondary row represents personal loan and column represents whether the customer is using online banking.

#B: Task to classify a customer who owns a bank Credit Card and is active online user, probability of customer accepting the loan.

#Setting predictors Personal Loan, CreditCard and Online

Mydata\_table <- train\_data[c("Personal Loan","CreditCard","Online")]  
Mydata\_table

## # A tibble: 3,000 × 3  
## `Personal Loan` CreditCard Online  
## <dbl> <dbl> <dbl>  
## 1 0 0 0  
## 2 0 0 0  
## 3 0 1 0  
## 4 0 0 1  
## 5 0 0 0  
## 6 0 0 1  
## 7 0 0 0  
## 8 1 0 0  
## 9 0 0 0  
## 10 1 0 0  
## # ℹ 2,990 more rows

#Create a pivot table with loan, Credit Card and online variables

Object1 <- ftable(Mydata\_table)  
Object1

## Online 0 1  
## Personal Loan CreditCard   
## 0 0 785 1145  
## 1 317 475  
## 1 0 65 122  
## 1 34 57

#Creating a pivot table only for credit card and Online variables.

Object2 <- ftable(Mydata\_table[,-1])  
Object2

## Online 0 1  
## CreditCard   
## 0 850 1267  
## 1 351 532

#Probability when Personal Loan = 1 , cc = 1 & online = 1

prob <- Object1[4,2]/Object2[2,2]   
prob

## [1] 0.1071429

#Insights : Using the above pivot tables, probability of customers who owns a credit card and is an active online banking user will accept loan. Number of customers with credit card = 1, Online = 1 and Personal Loan = 1 divided by total number of customers with credit card = 1 and Online = 1. The probability is 10.7% using training data set.

#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.

#Creating pivot table where Loan (rows) and Online (Columns)

Pivottable1 <- ftable(Mydata\_table[,-2])  
Pivottable1

## Online 0 1  
## Personal Loan   
## 0 1102 1620  
## 1 99 179

#Creating Pivot table where Loan(rows) and CC(Columns)

Pivottable2 <- ftable(Mydata\_table[,-3])  
Pivottable2

## CreditCard 0 1  
## Personal Loan   
## 0 1930 792  
## 1 187 91

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

#Total Observations

TotalCustomer <- nrow(Mydata\_table)

#Total number of customer accepted or rejected the loan (0 = NO, 1 = Yes)

Loanaccepted1 <- sum(Mydata\_table$`Personal Loan` == 1)  
Loanaccepted2 <- sum(Mydata\_table$`Personal Loan` == 0)

#Probability of credit card holder among the loan acceptors. P(CC = 1 | Loan = 1)

Probability1 <- sum(Mydata\_table$CreditCard == 1 & Mydata\_table$`Personal Loan` == 1)/Loanaccepted1  
Probability1

## [1] 0.3273381

#Probability of Online customers among the loan acceptors P(Online = 1 | Loan = 1)

Onlineprobability <- sum(Mydata\_table$Online == 1 & Mydata\_table$`Personal Loan` ==1)/Loanaccepted1  
Onlineprobability

## [1] 0.6438849

#Probability of loan acceptors P(Loan = 1)

Loanprobability <- Loanaccepted1 / TotalCustomer  
Loanprobability

## [1] 0.09266667

#Probability of credit card holder among the loan rejectors P(CC = 1 | Loan = 0)

Probability2 <- sum(Mydata\_table$CreditCard == 1 & Mydata\_table$`Personal Loan` == 0)/Loanaccepted2  
Probability2

## [1] 0.2909625

#Probability of Online customer among the loan rejectors P(Online = 1 | Loan = 0)

Onlineprobability1 <- sum(Mydata\_table$Online == 1 & Mydata\_table$`Personal Loan` ==0)/Loanaccepted2  
Onlineprobability1

## [1] 0.5951506

#Probability of loan rejectors P(Loan = 0)

Loanprobability1 <- Loanaccepted2 / TotalCustomer  
Loanprobability1

## [1] 0.9073333

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

#Computing Naive Bayes probability by manual calculation

numerator <- (Probability1 \* Onlineprobability \* Loanprobability) #(0.327 \* 0.643 \* 0.092)  
denominator <- (Probability1 \* Onlineprobability \* Loanprobability) + (Probability2 \* Onlineprobability1 \* Loanprobability1) #(0.327 \* 0.643 \* 0.092) + (0.290 \* 0.595 \* 0.907)  
  
nb\_probability <- numerator / denominator  
nb\_probability

## [1] 0.1105637

#Insights : Using the probabilities computed above, naive bayes isestimated. This formula assumes that Credit Card and Online areconditionally independent given loan. The probability estimated is 11.05%.

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

#Response Answer for F: The probability obtained from Pivot table(B) is 0.1071429 (10.7%) and Naive Bayes probability p(Loan = 1 | CC= 1, Online = 1) is 0.1105637(11.05%). The Pivot table probability is more accurate because it is directly based on observed counts from the training data set, reflecting the actual distribution of data. The Naive Bayes probability is an estimate value based on the assumption that variable Credit Card and Online are conditionally independent given Loan. This assumption may not hold properly in real data, leading to differences in result.

#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).

##Correcting the column name “Personal Loan” to Personal Loan and partitioning the data

set.seed(123)  
Mydata$PersonalLoan <- Mydata$`Personal Loan`  
Mydata <- Mydata %>% select(-'Personal Loan')  
train\_index <- sample(1:nrow(Mydata), 0.6 \* nrow(Mydata))  
train\_data <- Mydata[train\_index, ]  
validation\_data <- Mydata[-train\_index, ]

#Converting the categorical variables to factors ( Personal Loan,Credit Card, Online ) for training set

train\_data$CreditCard <- factor(train\_data$CreditCard, levels = c(0,1))  
train\_data$Online <- factor(train\_data$Online, levels = c(0,1))  
train\_data$PersonalLoan <- factor(train\_data$PersonalLoan, levels = c(0,1))

## Build Naive Bayes model on training data

nb\_model <- naive\_bayes(PersonalLoan ~ CreditCard + Online, data = train\_data)

#Creating a data frame for variables credit card and Online with factor levels (0,1)

prob\_train <- data.frame(  
 CreditCard = factor(1, levels = c(0,1)),  
 Online = factor(1, levels = c(0,1))  
)

#Predicting the probability for credit card and Online customers on training set

nb\_prob <- predict(nb\_model, prob\_train, type = "prob")  
nb\_prob

## 0 1  
## [1,] 0.8894363 0.1105637

#Converting the categorical variables to factors ( Personal Loan,Credit Card, Online ) for Validation set

validation\_data$CreditCard <- factor(validation\_data$CreditCard, levels = c(0,1))  
validation\_data$Online <- factor(validation\_data$Online, levels = c(0,1))  
validation\_data$PersonalLoan <- factor(validation\_data$PersonalLoan, levels = c(0,1))

#Creating a data frame for variables credit card and Online with factor levels (0,1)

prob\_valid <- data.frame(  
 CreditCard = factor(1, levels = c(0,1)),  
 Online = factor(1, levels = c(0,1))  
)

#Predicting the probability for credit card and Online customers on validation set

nb\_prob1 <- predict(nb\_model, prob\_valid, type = "prob")  
nb\_prob1

## 0 1  
## [1,] 0.8894363 0.1105637

#Insights : After apply the naive bayes classifier, the predicted probability of customer with CreditCard = 1, Online = 1 is 11.05% which matches with the manual calculation.