Assignment 3

MUNERAH

10/11/2021

```
library(readxl)
df<- read.csv("C:/Users/mnooo/Desktop/Datasets/UniversalBank.csv")
str(df)
```

```
## 'data.frame':    5000 obs. of  14 variables:
##  $ ID                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age               : int  25 45 39 35 35 37 53 50 35 34 ...
##  $ Experience        : int  1 19 15 9 8 13 27 24 10 9 ...
##  $ Income            : int  49 34 11 100 45 29 72 22 81 180 ...
##  $ ZIP.Code          : int  91107 90089 94720 94112 91330 92121 91711 93943 90089 93023 ...
##  $ Family            : int  4 3 1 1 4 4 2 1 3 1 ...
##  $ CCAvg             : num  1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...
##  $ Education         : int  1 1 1 2 2 2 2 3 2 3 ...
##  $ Mortgage          : int  0 0 0 0 0 155 0 0 104 0 ...
##  $ Personal.Loan     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Securities.Account: int  1 1 0 0 0 0 0 0 0 0 ...
##  $ CD.Account        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Online            : int  0 0 0 0 0 1 1 0 1 0 ...
##  $ CreditCard        : int  0 0 0 0 1 0 0 1 0 0 ...
```

```
#install.packages
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(class)
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```r
library(FNN)
```

```
##
## Attaching package: 'FNN'
```

```
## The following objects are masked from 'package:class':
##
##     knn, knn.cv
```

```r
library(e1071)
library(reshape2)
```

```r
### Change Numerical data to Catogerical
df$Personal.Loan<-factor(df$Personal.Loan)
df$Online<-factor(df$Online)
df$CreditCard <-factor(df$CreditCard)
str(df)
```

```
## 'data.frame':    5000 obs. of  14 variables:
##  $ ID                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age               : int  25 45 39 35 35 37 53 50 35 34 ...
##  $ Experience        : int  1 19 15 9 8 13 27 24 10 9 ...
##  $ Income            : int  49 34 11 100 45 29 72 22 81 180 ...
##  $ ZIP.Code          : int  91107 90089 94720 94112 91330 92121 91711 93943 90089 93023 ...
##  $ Family            : int  4 3 1 1 4 4 2 1 3 1 ...
##  $ CCAvg             : num  1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...
##  $ Education         : int  1 1 1 2 2 2 2 3 2 3 ...
##  $ Mortgage          : int  0 0 0 0 0 155 0 0 104 0 ...
##  $ Personal.Loan     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Securities.Account: int  1 1 0 0 0 0 0 0 0 0 ...
##  $ CD.Account        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Online            : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 2 1 ...
##  $ CreditCard        : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
```

```r
### Divide the data into 60% training and 40% validation
# First select the required variables
selected.var <- c(10,13,14)
set.seed(15)
bank.tr.in = createDataPartition(df$Personal.Loan,p=0.6, list=FALSE) # 60% reserved for Training
bank.tr = df[bank.tr.in,selected.var]
bank.va <- df[-bank.tr.in,selected.var] # Validation  data is rest
summary(bank.tr)
```

```
##  Personal.Loan Online   CreditCard
##  0:2712        0:1238   0:2128
##  1: 288        1:1762   1: 872
```

```
summary (bank.va)
```

```
##  Personal.Loan Online   CreditCard
##  0:1808        0: 778   0:1402
##  1: 192        1:1222   1: 598
```

```
###Create a pivot table for the training data with Online as a column variable, CC as a row vari
able, and Loan as a secondary row variable.

### using table function
table(bank.tr)
```

```
## , , CreditCard = 0
##
##              Online
## Personal.Loan    0    1
##             0  791 1130
##             1   82  125
##
## , , CreditCard = 1
##
##              Online
## Personal.Loan    0    1
##             0  330  461
##             1   35   46
```

```
##B. Consider the task of classifying a customer who owns a bank credit card and is actively usi
ng online banking services. Looking at the pivot table, what is the probability that this custom
er will accept the loan offer? [This is the probability of loan acceptance (Loan = 1) conditiona
l on having a bank credit card (CC = 1) and being an active user of online banking services (Onl
ine = 1)].


P1<-prop.table(table(bank.tr),margin = 2)
P1
```

```
## , , CreditCard = 0
##
##              Online
## Personal.Loan          0          1
##             0 0.63893376 0.64131669
##             1 0.06623586 0.07094211
##
## , , CreditCard = 1
##
##              Online
## Personal.Loan          0          1
##             0 0.26655897 0.26163451
##             1 0.02827141 0.02610670
```

Looking to the pivot table , There are 507 customers who owns a bank credit card and actively using online service , 46 of them will accept the loan The probability of a customer to accept the loan conditional on having a credit card and being an active of online services is (46/507)*100 = 9.07% In another way : 0.0261/(0.0261+0.261) = 0.09

```
##C. Create two separate pivot tables for the training data. One will have Loan (rows) as a func
tion of Online (columns) and the other will have Loan (rows) as a function of CC.

Pivot1 <- table(bank.tr$Personal.Loan,bank.tr$Online)
pivot_df1 <- as.data.frame(Pivot1)
colnames(pivot_df1) <- c("PersonalLoan", "Online")
pivot_df1
```

```
##   PersonalLoan Online   NA
## 1            0      0 1121
## 2            1      0  117
## 3            0      1 1591
## 4            1      1  171
```

```
Pivot2 <- table(bank.tr$Personal.Loan,bank.tr$CreditCard)
pivot_df2 <- as.data.frame(Pivot2)
colnames(pivot_df2) <- c("PersonalLoan", "CreditCard")
pivot_df2
```

```
##   PersonalLoan CreditCard   NA
## 1            0          0 1921
## 2            1          0  207
## 3            0          1  791
## 4            1          1   81
```

```
##D. Compute the following quantities [P(A | B) means “the probability of A given B

##i. P(CC = 1 | Loan = 1)

Pr1<- (table(bank.tr$CreditCard,bank.tr$Personal.Loan))
Pr1[2,2]/(Pr1[2,2]+Pr1[1,2])
```

```
## [1] 0.28125
```

```
Pr1
```

```
##
##        0    1
##   0 1921  207
##   1  791   81
```

The result as shown is 28%

```
##ii. P(Online = 1 | Loan = 1)

Pr2<-table(bank.tr$Online , bank.tr$Personal.Loan)
Pr2[2,2]/(Pr2[2,2]+Pr2[1,2])
```

```
## [1] 0.59375
```

```
Pr2
```

```
##
##        0    1
##   0 1121  117
##   1 1591  171
```

The result as shown 59%

```
##iii.P(Loan = 1) (the proportion of loan acceptors)
Pr3<-table(bank.tr$Personal.Loan)
Pr3[2]/(Pr3[2]+Pr3[1])
```

```
##     1
## 0.096
```

```
Pr3
```

```
##
##    0    1
## 2712  288
```

The result as shown approx 10%

```
##iv. P(CC = 1 | Loan = 0)
Pr4<-table(bank.tr$CreditCard,bank.tr$Personal.Loan)
Pr4[2,1]/(Pr4[2,1]+Pr4[1,1])
```

```
## [1] 0.2916667
```

```
Pr4
```

```
##
##        0    1
##   0 1921  207
##   1  791   81
```

The result as shown is 29%

```
##v.P(Online = 1 | Loan = 0)
Pr5<-table(bank.tr$Online , bank.tr$Personal.Loan)
Pr5[2,1]/(Pr4[2,1]+Pr4[1,1])
```

```
## [1] 0.5866519
```

```
Pr5
```

```
##
##        0    1
##   0 1121  117
##   1 1591  171
```

The result as shown is 59%

```
## vi. P(Loan = 0)
Pr6<-table(bank.tr$Personal.Loan)
Pr6[1]/(Pr5[1]+Pr5[2])
```

```
## 0
## 1
```

```
Pr6
```

```
##
##    0    1
## 2712  288
```

The result as shown approx 90%

```
##E. Use the quantities computed above to compute the naive Bayes probability P(Loan = 1 | CC =
 1, Online = 1)
### C1 loan , X1 CC ,X2 Online
###P(C1|X1,X2) =
##(P(X1|C1)P(X2|C1)P(C1) ) / P(X1|C1)P(X2|C1)P(C1)+P(X1|C2)*P(X2|C2)*P(C2)
x1<-((0.28125)*(0.593)*(0.096))
x2<-((0.28125)*(0.593)*(0.096))+((0.2916)*(0.5866)*(0.904))
nbresult<- (x1/x2)
nbresult
```

```
## [1] 0.09382773
```

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

# The result in B = 0.0907 ,in E =0.938,, there is no big different between them , and the value
which is calculated by pivot table is more accurate because the Naive Base assume the probabilit
ies being independent
```

```
##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).

bank.nb <- naiveBayes(Personal.Loan ~ ., data = bank.tr)
bank.nb
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     0     1
## 0.904 0.096
##
## Conditional probabilities:
##    Online
## Y           0         1
##   0 0.4133481 0.5866519
##   1 0.4062500 0.5937500
##
##    CreditCard
## Y           0         1
##   0 0.7083333 0.2916667
##   1 0.7187500 0.2812500
```

```{r}((0.28125)*(0.593)*(0.096)) / ((0.28125)*(0.593)*(0.096))+((0.2916)*(0.5866)*(0.904)) = 0.0938 it gives identical result as it is in the question E .

```