Assignment 3 Naive Bayes for Classification

Sasidhar Thatini

2023-11-04

#Step 1: Load and Explore the Data

```
# Load the necessary libraries
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
```

```
# Load the dataset
bank_data <- read.csv("C:\\Users\\T.Sasidhar\\Downloads\\UniversalBank-1.csv")

# Check the structure of the dataset
str(bank_data)
```

```
## '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 ...
```

```
# View the first few rows of the dataset
head(bank_data)
```

```
##   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
```

#Step 2: Partition the Data into Training and Validation Sets

```
set.seed(123)  # For reproducibility
n <- nrow(bank_data)
training_indices <- sample(1:n, 0.6 * n)  # 60% for training, 40% for validation
training_data <- bank_data[training_indices, ]
validation_data <- bank_data[-training_indices, ]
```

#Step 3: Create a Pivot Table for the Training Data

A. Create a pivot table using table() for the training data:

```
# Create a pivot table with conditional counts
pivot_table <- xtabs(~ CreditCard + Online + Personal.Loan, data = training_data)

# View the pivot table
pivot_table
```

```
## , , Personal.Loan = 0
##
##           Online
## CreditCard    0    1
##          0  785 1145
##          1  317  475
##
## , , Personal.Loan = 1
##
##           Online
## CreditCard    0    1
##          0   65  122
##          1   34   57
```

#Step 4: Calculate Probability P(Loan = 1 | CC = 1, Online = 1)

B. Calculate the probability of a customer accepting the loan offer given they have a credit card and use online banking:

```
# Probability of Loan acceptance (Loan = 1) conditional on CC = 1 and Online = 1
probability_loan_given_cc_online <- pivot_table[1, 1, 1] / sum(pivot_table[1, 1, ])
probability_loan_given_cc_online
```

```
## [1] 0.9235294
```

The output [1] 0.9235294 represents the probability that a customer who owns a bank credit card (CC = 1) and is actively using online banking services (Online = 1) will accept the loan offer (P(Loan = 1 | CC = 1, Online = 1)).

With a probability of approximately 0.9235 (or 92.35%), it suggests that there is a high likelihood that a customer who has a bank credit card and actively uses online banking services will accept the loan offer based on the Naive Bayes classifier's prediction.

This probability is significantly higher than the value obtained from the pivot table in Question B, which was approximately 10.74%. The Naive Bayes model seems to provide a much higher estimate of the probability of loan acceptance for customers with a credit card and online banking usage.

The conclusion for Question B is that, according to the Naive Bayes classifier, a customer who owns a bank credit card and is actively using online banking services is highly likely to accept the loan offer, with an estimated probability of approximately 92.35%. This estimate is substantially different from the pivot table value and may indicate that the Naive Bayes model captures different relationships in the data compared to the pivot table approach.

C. Create Two Separate Pivot Tables

#Create two separate pivot tables for Loan as a function of Online and Loan as a function of CC:

```r
# Pivot table for Loan as a function of Online
pivot_table_loan_online <- xtabs(~ Online + Personal.Loan, data = training_data)

# Pivot table for Loan as a function of CC
pivot_table_loan_cc <- xtabs(~ CreditCard + Personal.Loan, data = training_data)

# View the pivot tables
pivot_table_loan_online
```

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

```r
pivot_table_loan_cc
```

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

D. Compute the Following Quantities

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

```r
# Proportion of credit card holders among the loan acceptors
probability_cc_given_loan <- pivot_table_loan_cc[2, 2] / sum(pivot_table_loan_cc[2, ])
probability_cc_given_loan
```

```
## [1] 0.1030578
```

The output [1] 0.1030578 represents the probability that a customer who holds a bank credit card (CC = 1) did not accept a personal loan (Personal.Loan = 0), which is P(CC = 1 | Loan = 0). This probability indicates the proportion of credit card holders among customers who did not accept the personal loan.

With a probability of approximately 0.1031 (or 10.31%), it suggests that about 10.31% of customers who did not accept the personal loan also have a bank credit card.

The conclusion is that approximately 10.31% of customers who did not accept the personal loan also hold a bank credit card (CC = 1). This indicates that a relatively small proportion of non-loan acceptors have a bank credit card.

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

```r
# Proportion of online users among the loan acceptors
probability_online_given_loan <- pivot_table_loan_online[2, 2] / sum(pivot_table_loan_online[2, ])
probability_online_given_loan
```

```
## [1] 0.09949972
```

The output [1] 0.09949972 represents the probability that a customer who is actively using online banking services (Online = 1) did not accept a personal loan (Personal.Loan = 0), which is P(Online = 1 | Loan = 0). This probability indicates the proportion of online banking users among customers who did not accept the personal loan.

With a probability of approximately 0.0995 (or 9.95%), it suggests that about 9.95% of customers who did not accept the personal loan are active users of online banking services.

The conclusion is that approximately 9.95% of customers who did not accept the personal loan actively use online banking services (Online = 1). This indicates that a relatively small proportion of non-loan acceptors are active online banking users.

#iii. P(Loan = 1)

```r
# Proportion of loan acceptors
probability_loan <- sum(pivot_table_loan_online[2, ]) / nrow(training_data)
probability_loan
```

```
## [1] 0.5996667
```

The output [1] 0.5996667 represents the probability that a customer accepted a personal loan (Personal.Loan = 1), which is P(Loan = 1). This probability indicates the proportion of customers who accepted the personal loan among all the customers in the dataset.

With a probability of approximately 0.5997 (or 59.97%), it suggests that about 59.97% of customers in the dataset accepted the personal loan, while the remaining 40.03% did not accept the personal loan.

The conclusion is that approximately 59.97% of customers in the dataset accepted the personal loan (Loan = 1). This provides insight into the overall acceptance rate for personal loans among the customers in the dataset.

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

```r
# Proportion of credit card holders among those who did not accept the loan
probability_cc_given_no_loan <- pivot_table_loan_cc[2, 1] / sum(pivot_table_loan_cc[2, ])
probability_cc_given_no_loan
```

```
## [1] 0.8969422
```

The output [1] 0.8969422 represents the probability that a customer who holds a bank credit card (CC = 1) did not accept a personal loan (Personal.Loan = 0), which is P(CC = 1 | Loan = 0). This probability indicates the proportion of credit card holders among customers who did not accept the personal loan.

With a probability of approximately 0.8969 (or 89.69%), it suggests that a significant majority of customers who did not accept the personal loan also have a bank credit card (CC = 1).

The conclusion is that approximately 89.69% of customers who did not accept the personal loan have a bank credit card (CC = 1). This implies that a large proportion of non-loan acceptors are also credit card holders.

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

```r
# Proportion of online users among those who did not accept the loan
probability_online_given_no_loan <- pivot_table_loan_online[2, 1] / sum(pivot_table_loan_online[2, ])
probability_online_given_no_loan
```

```
## [1] 0.9005003
```

The output [1] 0.9005003 represents the probability that a customer who is actively using online banking services (Online = 1) did not accept a personal loan (Personal.Loan = 0), which is P(Online = 1 | Loan = 0). This probability indicates the proportion of online banking users among customers who did not accept the personal loan.

With a probability of approximately 0.9005 (or 90.05%), it suggests that a significant majority of customers who did not accept the personal loan are active users of online banking services (Online = 1).

The conclusion is that approximately 90.05% of customers who did not accept the personal loan actively use online banking services (Online = 1). This implies that a large proportion of non-loan acceptors are also active online banking users.

#vi. P(Loan = 0)

```
# Proportion of non-loan acceptors
probability_no_loan <- 1 - probability_loan
probability_no_loan
```

```
## [1] 0.4003333
```

The output [1] 0.4003333 represents the probability that a customer did not accept a personal loan (Personal.Loan = 0), which is P(Loan = 0). This probability indicates the proportion of customers who did not accept the personal loan among all the customers in the dataset.

With a probability of approximately 0.4003 (or 40.03%), it suggests that about 40.03% of customers in the dataset did not accept the personal loan, while the remaining 59.97% accepted the personal loan.

The conclusion is that approximately 40.03% of customers in the dataset did not accept the personal loan (Loan = 0). This provides insight into the overall non-acceptance rate for personal loans among the customers in the dataset.

E. Use the Quantities to Compute Naive Bayes Probability P(Loan = 1 | CC = 1, Online = 1)

```
# Use Naive Bayes formula to calculate P(Loan = 1 | CC = 1, Online = 1)
naive_bayes_probability <- (probability_loan_given_cc_online * probability_loan) / (probability_cc_given
naive_bayes_probability
```

```
## [1] 54.008
```

The output [1] 54.008 represents the Naive Bayes probability that a customer who holds a bank credit card (CC = 1) and is actively using online banking services (Online = 1) will accept the loan offer (P(Loan = 1 | CC = 1, Online = 1)).

With a probability of approximately 54.008 (or 5400.8%), it suggests that, according to the Naive Bayes classifier, there is a very high likelihood that a customer who holds a bank credit card (CC = 1) and is actively using online banking services (Online = 1) will accept the loan offer.

According to the Naive Bayes classifier, the estimated probability of a customer accepting the loan offer when they have both a bank credit card and are active online banking users is approximately 5400.8%. This is a significantly high probability estimate.

#F. Compare the Naive Bayes Probability with the Pivot Table Value

- For Question E: P(Loan = 1 | CC = 1, Online = 1) is approximately 54.008%.
- For Question B: The probability that a customer who owns a bank credit card and is actively using online banking services will accept the loan offer is approximately 92.35%.

To compare these values and determine which is a more accurate estimate, we need to consider the context:

Question E's estimate (54.008%) is obtained from a Naive Bayes model that takes into account the conditional probabilities of CreditCard and Online usage given the target variable Personal.Loan. It's a model-based estimate.

Question B's estimate (92.35%) is based on direct observation of the data, specifically from the pivot table that shows the count of customers with specific combinations of CreditCard and Online usage who accepted the loan offer. It's a frequency-based estimate.

In this case, the estimate from Question B (pivot table) is likely a more accurate reflection of the data because it's directly derived from the observed counts in the dataset. It provides an empirical probability based on the actual data, while the Naive Bayes model in Question E makes certain assumptions and approximations.

Therefore, for this specific scenario, the estimate from Question B is considered more accurate because it is grounded in the data itself. The Naive Bayes model in Question E may have limitations or assumptions that lead to differences in the estimated probabilities.

#Step G: Run Naive Bayes on the Data

#To run Naive Bayes on the data, we can use the naiveBayes function from the e1071 package.

```
# Install and load the e1071 package
install.packages("e1071")
```

```
## Installing package into 'C:/Users/T.Sasidhar/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'e1071' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'e1071'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\T.Sasidhar\AppData\Local\R\win-library\4.3\00LOCK\e1071\libs\x64\e1071.dll
## to
## C:\Users\T.Sasidhar\AppData\Local\R\win-library\4.3\e1071\libs\x64\e1071.dll:
## Permission denied

## Warning: restored 'e1071'

##
## The downloaded binary packages are in
##  C:\Users\T.Sasidhar\AppData\Local\Temp\RtmpEn81z9\downloaded_packages
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.3.2
```

```
# Create a data frame with only the relevant variables (CreditCard, Online, Personal.Loan)
data_for_naive_bayes <- training_data[, c("CreditCard", "Online", "Personal.Loan")]

# Fit a Naive Bayes model
naive_bayes_model <- naiveBayes(Personal.Loan ~ CreditCard + Online, data = data_for_naive_bayes)

# View the model output
naive_bayes_model
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##          0          1
## 0.90733333 0.09266667
```

```
##
## Conditional probabilities:
##    CreditCard
## Y          [,1]      [,2]
##   0 0.2909625 0.4542897
##   1 0.3273381 0.4700881
##
##    Online
## Y          [,1]      [,2]
##   0 0.5951506 0.4909531
##   1 0.6438849 0.4797134
```

In Question G, we are asked to compare the Naive Bayes model's output to the number obtained in Question E and identify the specific entry in the model's output that corresponds to P(Loan = 1 | CC = 1, Online = 1). Let's examine the outputs and draw a conclusion based on the provided information:

Output for Question E: P(Loan = 1 | CC = 1, Online = 1) = 54.008

Output for Naive Bayes Model: The Naive Bayes model provides conditional probabilities for CreditCard and Online based on the target variable Personal.Loan (Loan). Here's a summary of the relevant conditional probabilities from the model's output:

Conditional probabilities for CreditCard: - P(CC = 1 | Loan = 0) = 0.3273381 - P(CC = 1 | Loan = 1) = 0.4700881

Conditional probabilities for Online: - P(Online = 1 | Loan = 0) = 0.6438849 - P(Online = 1 | Loan = 1) = 0.4797134

To compute P(Loan = 1 | CC = 1, Online = 1), you can use the conditional probabilities for CC and Online when Loan = 1:

P(CC = 1 | Loan = 1) = 0.4700881 P(Online = 1 | Loan = 1) = 0.4797134

Now, calculate P(Loan = 1 | CC = 1, Online = 1) using these conditional probabilities:

P(Loan = 1 | CC = 1, Online = 1) = P(CC = 1 | Loan = 1) * P(Online = 1 | Loan = 1) P(Loan = 1 | CC = 1, Online = 1) 0.4700881 * 0.4797134-0.22565 (approximately 22.57%)

Comparison and Conclusion: The Naive Bayes model's estimated probability P(Loan = 1 | CC = 1, Online = 1) is approximately 22.57%. This is significantly different from the value obtained in Question E, which was 54.008%.

The conclusion is that the Naive Bayes model provides a lower estimate for P(Loan = 1 | CC = 1, Online = 1) compared to the value obtained in Question E. This suggests that the model's prediction differs from the direct computation and may indicate that the model captures different relationships or has potential limitations in representing the data accurately. It's important to consider the model's output in the context of its assumptions and the data used.