Assignment32

John Bute

2024-11-15

UniversalBank is looking at converting its liability customers (i.e., customers who only have deposits at the bank) into asset customers (i.e., customers who have a loan with the bank). In a previous campaign, UniversalBank was able to convert 9.6% of 5000 of its liability customers into asset customers. The marketing department would like to understand what combination of factors make a customer more likely to accept a personal loan, in order to better design the next conversion campaign. UniversalBank.csv contains data on 5000 customers, including the following measurements: age, years of professional experience, yearly income (in thousands of USD), family size, value of mortgage with the bank, whether the client has a certificate of deposit with the bank, a credit card, etc. They build 2 decision trees on a training subset of 3000 records to predict whether a customer is likely to accept a personal loan (1) or not (0).

a) Explore UniversalBank.csv. Can you come up with a reasonable guess as to what each of the variables represent?

```
bank_data <- read.csv("C:/Users/johnb/Desktop/Machine
Learning/data/UniversalBank.csv")

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
```

ID: ID of the customer Age: age of the customer Experience: Years of professional experience Income: Income in thousands Zip Code: zip code of the client Family: number of family members including the customer CCAvg: Average credit card spending in the thousands Education: encoded? possibly 1 = Undergrad 2 = Graduate Mortgage: Mortgage in the customer's name (thousands) Personal.Loand: whether the customer has a personal loan Securities.Account: wherther he has a securities account CD.Account: does the customer have a Certificate of deposit account Online: does the customer use online banking CreditCard: does the customer have a credit card.

b)How many variables are used in the construction of tree *A*? Of tree *B*? Tree A uses 5 variables, while tree B uses only 2.

c) Are the following decision rules valid or not for trees *A* and/or *B*? IF (Income ≥ 114) AND (Education ≥ 1.5) THEN (Personal Loan = 1) IF (Income < 92) AND (CCAvg ≥ 3) AND (CD.Account < 0.5) THEN (Personal Loan = 0)

The first statement is valid for both tree A and B The second statement is valid for tree A and B

d) What prediction would trees *A* and *B* make for a customer with: a yearly income of 94,000$USD (Income = 94); 2 kids (Family = 4); no certificate of deposit with the bank (CD.Account = 0); a credit card interest rate of 3.2% (CCAvg = 3.2), and a graduate degree in Engineering (Education = 3)?

Tree A will predict that yes, they do have a personal loan. Meanwhile, tree B willpredict that they do not have one.

9. The confusion matrices for the predictions of trees *A* and *B* on the remaining 2000 testing observations are shown below.

Using the appropriate matrices, compute the 9 performance evaluation metrics for each of the trees (on the testing set).

The 9 performance evaluation metrics are:

Accuracy, Precision, Recall, F1 Score, False Positive Rate, False Negative Rate, Negative Predictive Value, Balanced Accuracy.

```r
confusion_matrix_A <- matrix(c(1792, 18, 19, 171), nrow = 2, byrow = TRUE)

rownames(confusion_matrix_A) <- c("Actual_A", "Actual_B")
colnames(confusion_matrix_A) <- c("Pred_A", "Pred_B")


confusion_matrix_B <- matrix(c(1801, 64, 10, 125), nrow = 2, byrow = TRUE)

rownames(confusion_matrix_B) <- c("Actual_A", "Actual_B")
colnames(confusion_matrix_B) <- c("Pred_A", "Pred_B")

compute_metrix <- function(matrix) {

  TP <- matrix[1,1]
  TN <- matrix[2,2]
  FP <- matrix[1,2]
  FN <- matrix[2,1]
  accuracy <- (TP + TN) / sum(matrix)

  precision <- TP / (TP + FP)

  recall <-TP / (TP + FN)

  specificity <- TN / (TN + FP)

  f1_score <- 2 * (precision * recall) / (precision + recall)

  FPR <- FP / (FP + TN)

  FNR <- FN / (TP + FN)

  NPV <- TN / (TN + FN)

  balanced_accuracy <- (recall + specificity) / 2

  data.frame(
    Accuracy = accuracy,
    Precision = precision,
    Recall = recall,
    Specificity = specificity,
    f1_score = f1_score,
    False_Positive_Rate = FPR,
    False_Negative_Rate = FNR,
    Negative_Predictive_Value = NPV,
    Balanced_Accuracy = balanced_accuracy
  )

}

metrics_A <- compute_metrix(confusion_matrix_A)

metrics_B <- compute_metrix(confusion_matrix_B)

print(metrics_A)
```

```
##   Accuracy Precision    Recall Specificity  f1_score False_Positive_Rate
## 1   0.9815 0.9900552 0.9895086   0.9047619 0.9897818           0.0952381
##   False_Negative_Rate Negative_Predictive_Value Balanced_Accuracy
## 1          0.01049144                       0.9         0.9471352
```

```
print(metrics_B)
```

```
##   Accuracy Precision    Recall Specificity  f1_score False_Positive_Rate
## 1    0.963 0.9656836 0.9944782   0.6613757 0.9798694           0.3386243
##   False_Negative_Rate Negative_Predictive_Value Balanced_Accuracy
## 1         0.005521811                 0.9259259         0.8279269
```

b) If customers who would not accept a personal loan get irritated when offered a personal loan, what tree should UniversalBank's marketing group use to help maintain good customer relations?

This scenario corresponds to a false positive, where the actual outcome is negative (no personal loan wanted) but the predicted outcome was positive. Therefore we need to look at the false positive rate and see which one is lower. For tree A, this value is 0.0952, while for tree B it is 0.3386. Thus since tree A's FPR is significantly lower than tree B, UniversalBank's marketing groups should use tree A.