Q32

Rahul Atre

2023-11-14

Q32 [Classification]

Do exercises 8 and 9, from DUDADS, chapter 19.

Exercise 8: 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 the UniversalBank.csv dataset. Can you come up with a reasonable guess as to what each of the variables represent?

Ans: First, let us load this dataset into an R dataframe

```
bank_df = read.csv("UniversalBank.csv")
```

In RStudio's environment tab, it says that the data contains 5000 obs. and 14 variables. Here is an intuitive guess as to what each of the variables represent:

- ID: Identification of the customer
- Age: Age of the customer (How old they are)
- Experience: Years of experience (How long they have worked)
- Income: The overall income of the customer per year in thousands
- ZIP.code: The ZIP Code (Address) of where the customer lives
- Family: Number of members in the customer's family
- CCAvg: Average amount spent on credit card per month in thousands
- Education: A logistic indicator of the Customer's educational level (high school, bachelors, masters, PHD, etc)
- Mortgage: Total amount of money borrowed from bank for house mortgage in thousands
- Personal.Loan: Binary indicator of whether the customer has taken a personal loan from bank or not
- Securities.Account: Binary indicator of whether the customer has a securities account with the bank
- CD.Account: Binary indicator of whether the customer has a cerificate of deposit account with the bank

- Online: Binary indicator of whether the customer registered online or not
- Credit.Card: Binary indicator of whether the customer has a credit card or not

b. How many variables are used in the construction of tree A? Of tree B?

- For the construction of tree A, 5 unique variables are used
- For the construction of tree B, 2 variables are used

c. Are the following decision rules valid or not for trees A and/or B?

- IF(Income $\geq$ 114) AND (Education $\geq$ 1.5) THEN (Personal Loan = 1)
- IF(Income $<$ 92) AND (CCAvg $\geq$ 3) AND (CD.Account $<$ 0.5) THEN (Personal Loan = 0)

For the following decision rules, we will trace the events individually and see if we get the correct outcome for tree A and B

For tree A: i) - IF(Income $\geq$ 114) -> No (Right) - (Education $\geq$ 1.5) -> No (Right) - Outcome: Personal Loan = 1 **Valid**

ii)

- (Income $<$ 92) -> Yes (Left)
- (CCAvg $\geq$ 3) -> No (Right)
- (CD.Account $<$ 0.5) -> Yes (Left)
- (Income $<$ 92) -> Yes (Left)
- Outcome: Personal Loan = 0 **[Valid]**

For tree B: i) - IF(Income $\geq$ 114) -> No (Right) - (Education $\geq$ 1.5) -> No (Right) - Outcome: Personal Loan = 1 **[Valid]**

ii)

- IF(Income $<$ 92) -> Yes (Left)
- Outcome: Personal Loan = 0 **[Valid]** (We don't need to check the other events since the outcome for the tree has already been decided)

Therefore, both decision rules are valid for both trees A and B.

d. What predictions 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 would make a prediction of Personal Loan = 0 (i.e. no loan granted), whereas Tree B would make a prediction of Personal Loan = 0 (i.e. no loan granted). Both tree's come to the same decision of not granting the customer a personal loan.

Exercise 9: The confusion matrices for the predictions of trees A and B on the remaining 2000 testing observations are shown below. (Image provided in textbook)

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

Let us create a function that automates the computation of all performance evaluation metrics that are given in the textbook:

```
performance_metrics_func<- function(TP, FN, AP, FP, TN, AN, PP, PN, t) {
  sensitivity <- paste("Sensitivity: ", TP/AP)
  specificity <- paste("Specificity: ", TN/AN)
  precision <- paste("Precision: " ,  TP/PP)
  neg_predictive_value <- paste("Negative Predictive Value: ", TN/PN)
  false_positive_rate <- paste("False Positive Rate: ", FP/AN)
  false_discovery_rate <- paste("False Discovery Rate: ", 1 - TP/PP)
  false_negative_rate <- paste("False Negative Rate: ", FN/AP)
  accuracy <- paste("Accuracy: " , (TP + TN)/t)
  f1_score <- paste("F1-score: ", 2*TP/(2*TP + FP + FN))
  MCC <- paste("MCC: ", (TP*TN - FP*FN)/(sqrt(AP*AN*PP*PN)))
  informedness_ROC <- paste("Informedness/ROC: ",  TP/AP + TN/AN - 1)
  markedness <- paste("Markedness: ", TP/PP + TN/PN - 1)

  output <- cat(sensitivity, specificity, precision, neg_predictive_value, false_positive_rate, false_di
}
```

Tree A performance metrics:

```
#Tree A
performance_metrics_func(1792, 19, 1811, 18, 171, 189, 1810, 190, 2000)
```

```
## Sensitivity:  0.989508558807289
## Specificity:  0.904761904761905
## Precision:  0.990055248618785
## Negative Predictive Value:  0.9
## False Positive Rate:  0.0952380952380952
## False Discovery Rate:  0.00994475138121542
## False Negative Rate:  0.0104914411927112
## Accuracy:  0.9815
## F1-score:  0.989781828224247
## MCC:  0.892160366629517
## Informedness/ROC:  0.894270463569194
## Markedness:  0.890055248618784
```

Tree B performance metrics:

```
#Tree B
performance_metrics_func(1801, 10, 1811, 64, 125, 189, 1865, 135, 2000)
```

```
## Sensitivity:  0.994478188845942
## Specificity:  0.661375661375661
## Precision:  0.9656836461126
## Negative Predictive Value:  0.925925925925926
## False Positive Rate:  0.338624338624339
## False Discovery Rate:  0.0343163538873995
```

```
## False Negative Rate:  0.00552181115405853
## Accuracy:  0.963
## F1-score:  0.979869423286181
## MCC:  0.764699660465403
## Informedness/ROC:  0.655853850221603
## Markedness:  0.891609572038526
```

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?

If this customer would get irritated from an inaccurate prediction, then the metrics of accuracy and false positive rate would be most important:

Tree A - Accuracy: 0.9815 = 98.15% - False Positive Rate: 0.09523 = 9.523%

Tree B - Accuracy: 0.963 = 96.3% - False Positive Rate: 0.3386 = 33.86%

Tree A has a higher accuracy and lower false positive rate. Therefore, The tree that UniversalBank's marketing group should use to maintain good customer relations is **Tree A**.