7.2 Personal Loan Acceptance

Universal Bank is a relatively young bank growing rapidly in terms of overall customer acquisition. The majority of these customers are liability customers (depositors) with varying sizes of relationship with the bank. The customer base of asset customers (borrowers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business. In particular, it wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).

A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise smarter campaigns with better target marketing. The goal is to use k-NN to predict whether a new customer will accept a loan offer. This will serve as the basis for the design of a new campaign.

The file UniversalBank.csv contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer’s relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (=9.6%) accepted the personal loan that was offered to them in the earlier campaign.

Partition the data into training (60%) and validation (40%) sets.

1. Consider the following customer:

|  |  |  |  |
| --- | --- | --- | --- |
| Age = 40 | Experience = 10 | Income = 84 | Family = 2 |
| CCAvg = 2 | Education\_1 = 0 | Education\_2 = 1 | Education\_3 = 0 |
| Mortgage = 0 | Securities Account = 0 | CD Account = 0 | Online = 1 |
| Credit Card = 1 |  |  |  |

Perform a k-NN classification with all predictors except ID and ZIP code using k = 1. Remember to transform categorical predictors with more than two categories into dummy variables first. Specify the success class as 1 (loan acceptance), and the default cutoff of 0.5. How would this customer be classified?

1. What is a choice of k that balances between overfitting and ignoring the predictor information?
2. Show the confusion matrix for the validation data that results from using the best k.
3. Consider the following customer:

|  |  |  |  |
| --- | --- | --- | --- |
| Age = 40 | Experience = 10 | Income = 84 | Family = 2 |
| CCAvg = 2 | Education\_1 = 0 | Education\_2 = 1 | Education\_3 = 0 |
| Mortgage = 0 | Securities Account = 0 | CD Account = 0 | Online = 1 |
| Credit Card = 1 |  |  |  |

Classify the customer using the best k.

1. Repartition the data, this time into training, validation, and test sets (50%, 30%, 20%). Apply the k-NN method with the k chosen above. Compare the confusion matrix of the test set with that of the training and validation sets. Comment on the differences and their reason.
2. Please redo HW4 Q2 using the Random Forest method (other requirements do not change)

Q2. Please download **spambase.csv** from Beachboard, and build

1. A decision tree classifier
   1. *As demonstrated in class on October 12, 2021, the graphics method produces an error in VSCode. Therefore, a screenshot of the Confusion Table is presented instead.*

Requirement:

1. Using 30% of the data as testing data
2. Display the confusion matrix of both training data and testing data
3. Try to find a good tree with reasonable parameters and performance. Do not just give me a full tree!

This data description is in the file spambaseDOCUMENTATION.txt