Assignment Instructions: Assignment 2

Purpose

The purpose of this assignment is to use k-NN for classification.

Directions

Universal bank is a 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 in more loan business. In particular, it wants to explore ways of converting its liability customers to personal loan customers.

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.

Consider the following customer:

1. 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, and 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 use the default cutoff value of 0.5. How would this customer be classified?
2. What is a choice of k that balances between overfitting and ignoring the predictor information?
3. Show the confusion matrix for the validation data that results from using the best k.
4. 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 and Credit Card = 1. Classify the customer using the best k.

5. 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.

File Attached: UniversalBank.csv

Learning Outcomes

The assignment will help you with the following course outcomes:

1. Think critically about how to use machine learning algorithms to solve a given business problem.
2. Know how to formulate business problems and identify relevant data to use in modeling frameworks.
3. Know how to evaluate the appropriateness and estimate the performance of using k-NN for a given task.
4. Know how to use software tools (such as R) effectively to implement k-NN.
5. Foster the communication and presentation of statistical results and inferences.

Requirements

All due dates are included in the Assignment Schedule.

General Submission Instructions

All work must be your own. Copying other people's work or from the Internet is a form of plagiarism and will be prosecuted as such.

1. Create a new folder called **Assignment_2** in your previously created GitHub repository.
2. If you are using R, then upload the R Markdown file, the knitted pdf/html file, and any other data file you might have used for the assignment.
3. If you using Python, then share the Jupyter/Google Colab notebook in our Assignment_2 folder on GitHub

Provide the link to your git repository in Canvas for the assignment.