HW2

February 1, 2024

Problem 2.1 (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.

a. 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`. 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). How would this customer be classified?

b. What is a choice of k that balances between overfitting and ignoring the predictor information?

c. 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 obtained in Part b.

Problem 2.2

Return to the Universal Bank example described above, where the bank's goal is to identify new customers most likely to accept a personal loan.

Partition the data (60% training and 40% validation) and then perform a discriminant analysis that models Personal Loan as a function of the remaining predictors (excluding `Zip Code`). Remember to turn categorical predictors with more than two categories into dummy variables first. Specify the success class as 1 (personal loan acceptance), and use the default cutoff value of 0.5.

a. Compute summary statistics for the predictors separately for loan acceptors and nonacceptors. For continuous predictors, compute the mean and standard deviation. For categorical predictors, compute the percentages. Are there predictors where the two classes differ substantially?

b. Examine the model performance on the validation set. What is the accuracy rate? Is one type of misclassification more likely than the other?

c. (*optional*) Select three customers who were misclassified as acceptors and three who were misclassified as nonacceptors. The goal is to determine why they are misclassified. First, examine their probability of being classified as acceptors: is it close to the threshold of 0.5? If not, compare their predictor values to the summary statistics of the two classes to determine why they were misclassified.