Machine learning assignment 1

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.

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, 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?

Response:

According to the KNN output, this customer would be classified as accepting the personal loan offer with parameter k held constant at 1. The accuracy is 95.6%.

```
[1] 0.2868566
k-Nearest Neighbors

Resampling results:

  Accuracy   Kappa
  0.9559683  0.7064722
```

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

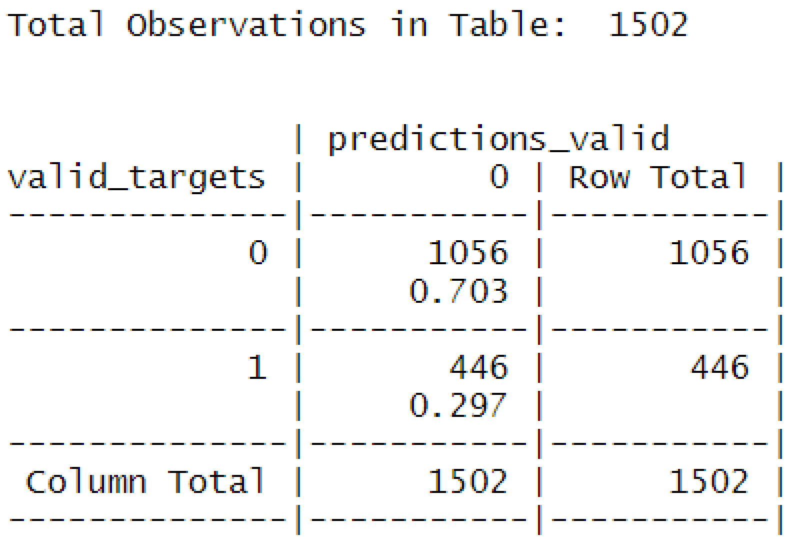
Response:

A choice of K that balances between over-fitting and ignoring the predictor information is K = 5.

```
k   Accuracy   Kappa
   5  0.9556791  0.6852176
```

3. Show the confusion matrix for the validation data that results from using the best k.

Response:



```
 Total Observations in Table:  1502 

 
               | predictions_valid 
 valid_targets |         0 | Row Total | 
---------------|-----------|-----------|
             0 |      1056 |      1056 | 
               |     0.703 |           | 
---------------|-----------|-----------|
             1 |       446 |       446 | 
               |     0.297 |           | 
---------------|-----------|-----------|
  Column Total |      1502 |      1502 | 
---------------|-----------|-----------|
```

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.

Response:

Using the best k, the customer would accept the loan, so he/she is in the loan acceptance class.

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.

Response:

The model performs better on the training data for the best k of 5. Although, for the test and validation data the model ability to predict is very poor. Meaning that the ability of the model to perform in predicting the target is significantly reduced from the model’s performance on the training data. Please see RMD file for code and full r output for this responses.