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. Specify the success class as 1 (loan acceptance), and use the default cutoff value of 0.5. How would this customer be classified?

Ans: **Accuracy k 1**
0.983017

With default as 0.5 the customer is classified as 0 0 0 (Deny Deny Deny) as no loan acceptance.

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

Ans: **Training and Validation**

k	**accuracy**
4	**0.9584792**

<u>For K=4 its 95% accuracy so 4is the best choice for the best fit model.</u>

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

Ans **For K=4**
confusionMatrix(nn4,vlddata$Personal.Loan)$overall[1]
Accuracy
0.9584792

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.

Ans : Customer with the best **K=4**

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.

Ans:

K=4
# Train and test
Accuracy
0.9661355
#test and valid
Accuracy
0.9594392
Difference 0.0066963 is observed due to cross validation.

****