Assignment #1

Deepak-George Thomas

Course – IE 583

Note – In this dataset it is tempting to remove the variable own_telephone. However, many credit card companies may use this criterion to judge how contactable the client is. Therefore, this variable is kept.

At the same time, it is evident that not all variables/attributes have equal importance. The Naïve Bayes classifier won't be realistic in this case because not all attributes can be assigned equal importance. kNN usually works best for numerical attributes. This dataset contains many qualitative attributes. It is expected that decision tree will give the best predictions as the assumption of attributes having a hierarchy is very well satisfied.

1. Decision Tree

In order to obtain the true error, both independent data sets as well as cross validation was used. Since this is a relatively small data set, we expect that CV to have better applicability than independent test set.

1.1 Independent Test Data

In order to account for variance for different confusion matrices were developed.

1.1.1 Test error

 # Repetition = 1 (set.seed(123))

```
Prediction bad good
      bad   35   25
      good  64  206

               Accuracy : 0.7303
                 95% CI : (0.679, 0.7774)
```

#Repetition = 2 (set.seed(1234))

```
Prediction bad good
      bad   32   31
      good  67  200

               Accuracy : 0.703
                 95% CI : (0.6505, 0.7518)
```

#Repetition = 3 (set.seed(12345))

```
Confusion Matrix and Statistics

          Reference
Prediction bad good
      bad   40   48
      good  59  183
```

```
               Accuracy : 0.6758
                 95% CI : (0.6223, 0.726)
```

#Repetition = 4 (set.seed(12))

```
Confusion Matrix and Statistics

          Reference
Prediction bad good
      bad   49   34
      good  50  197

               Accuracy : 0.7455
                 95% CI : (0.6949, 0.7916)
```

The mean error is .29 with a standard deviation of 0.02

1.1.2 Training error

 # Repetition = 1 (set.seed(123))

```
Prediction bad good
      bad  112   16
      good  89  453

               Accuracy : 0.8433
                 95% CI : (0.8135, 0.87)
```

 # Repetition = 2 (set.seed(1234))

```
Prediction bad good
      bad  127   15
      good  74  454

               Accuracy : 0.8672
                 95% CI : (0.8391, 0.8919)
```

 # Repetition = 3 (set.seed(12345))

```
Prediction bad good
      bad  140   29
      good  61  440

               Accuracy : 0.8657
                 95% CI : (0.8375, 0.8906)
```

#Repetition = 4 (set.seed(12))

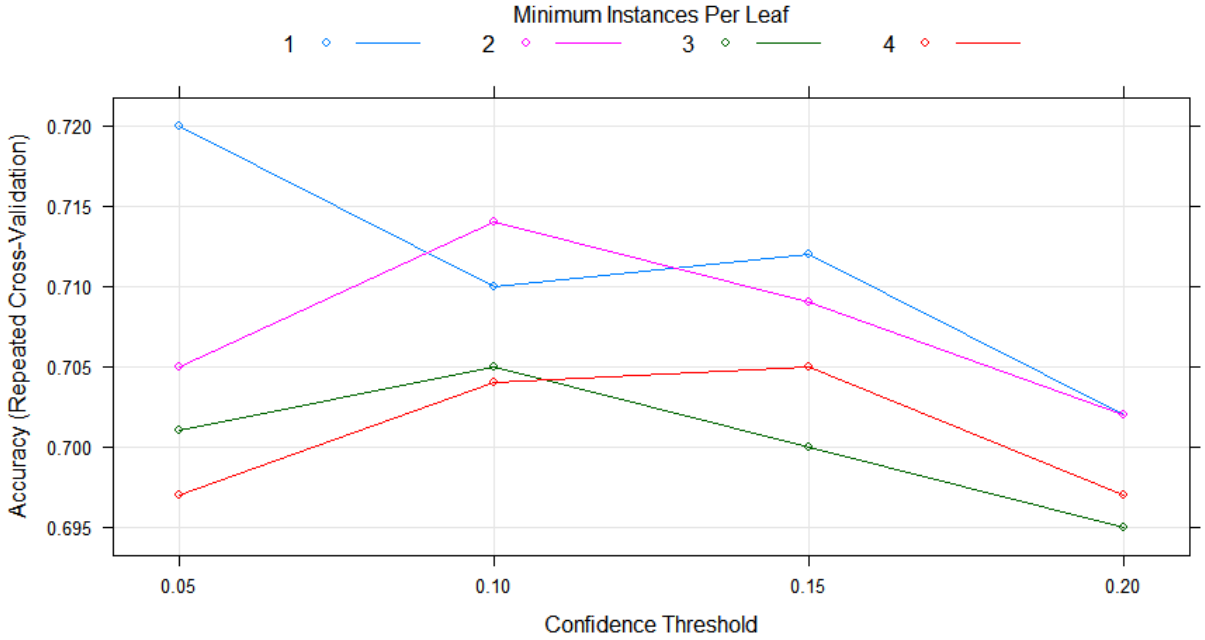
```
Prediction bad good
      bad  133   27
      good  68  442

               Accuracy : 0.8582
                 95% CI : (0.8295, 0.8837)
```

The mean error here is 0.14 with a standard deviation of 0.009
As expected, the training error is smaller than the test error. However the t raining error won't be used as it is not a reliable estimate

1.2 Cross Validatation

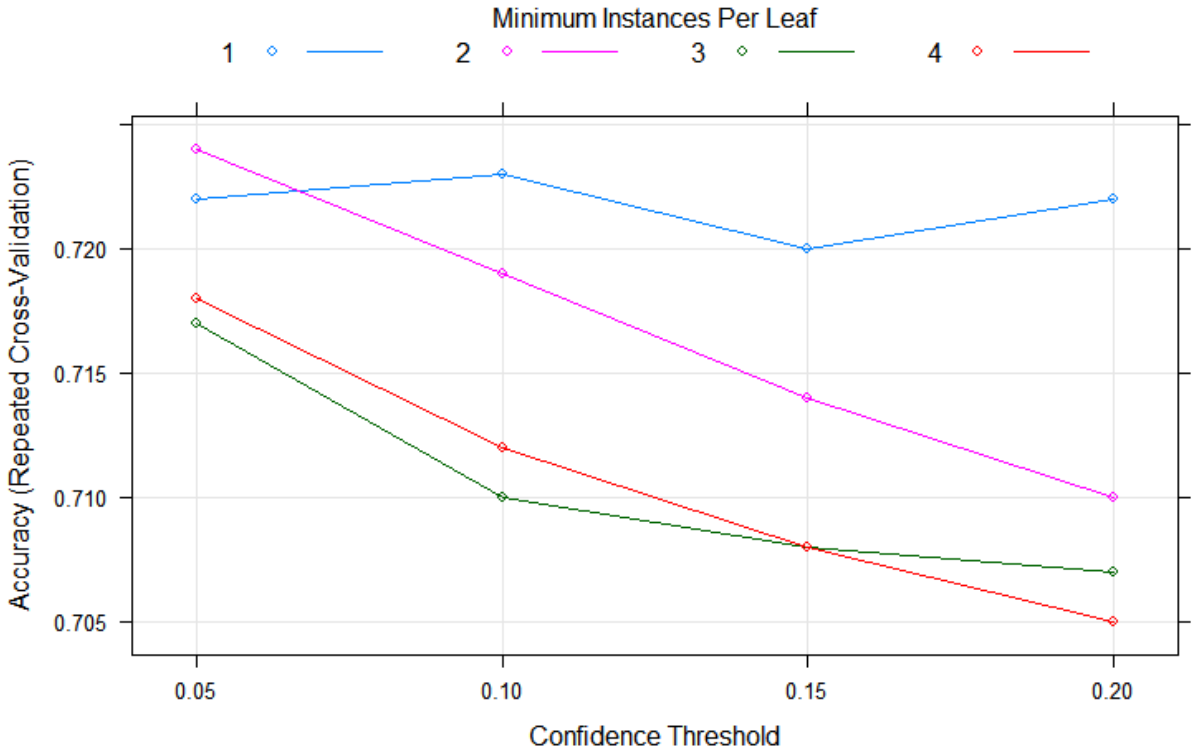
In order to tune our model, the minimum number of instances in a leaf (M) and the confidence threshold (C) were varied. M ranged from 1 to 4 and C r aranged from 0.05 to 0.20. The plot below describes the variation of accur acy with change in C & M.

1.2.1 Number of folds = 10



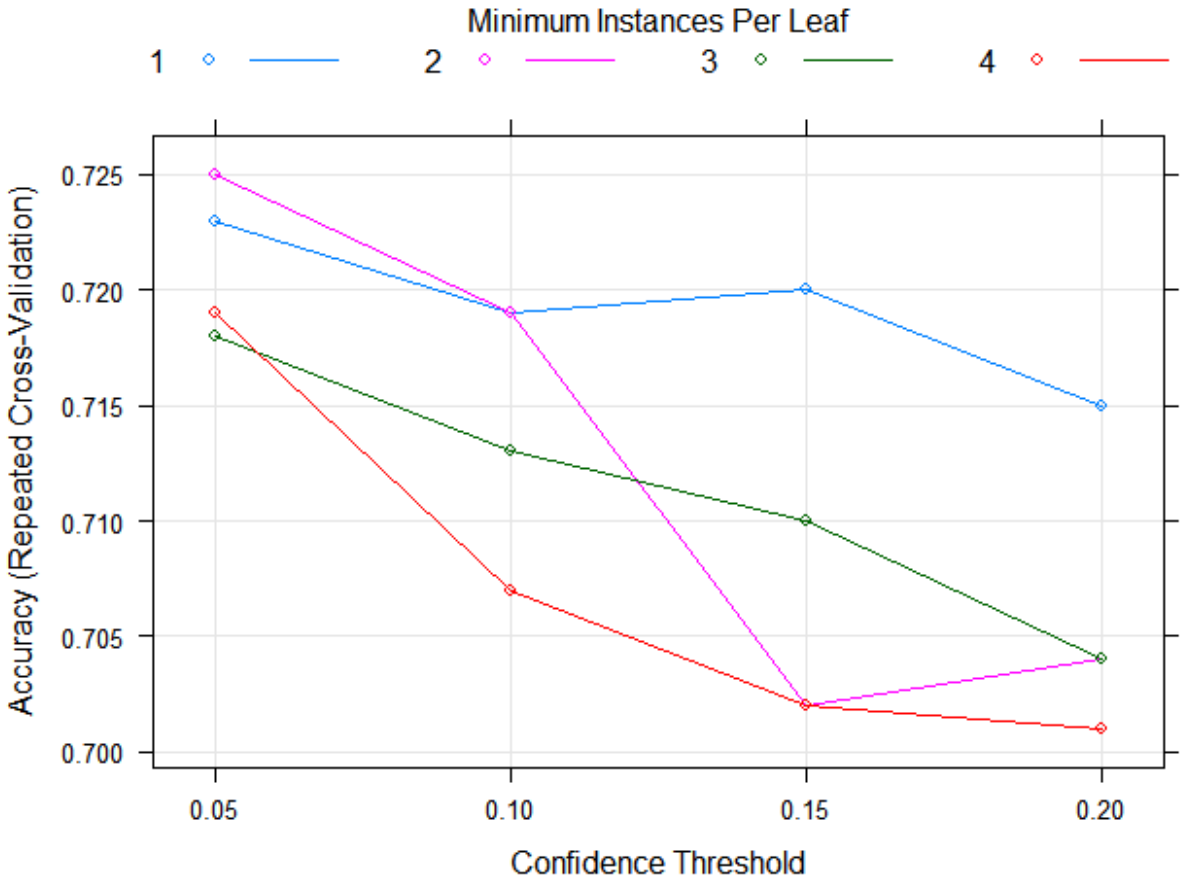
It is evident that the best accuracy of 0.720 is obtained when C = 0.05 and M = 1.

1.2.2 Number of folds = 5



It is evident that the best accuracy of 0.724 is obtained when C = 0.05 and M = 2.

1.2.3 Number of folds = 20



It is evident that the best accuracy of 0.725 is obtained when C = 0.05 and M = 2.

The average error estimate is coming out to be 0.28 with a standard deviation of 0.002. The error estimate obtained from CV is much larger than that obtained from test error with a much lower variability as expected.

KNN

1.3 Independent Test Data

2.1.1 Test error

```
#Repetition 1 (set.seed(123456))
Prediction bad good
      bad   23   46
      good  76  185

               Accuracy : 0.6303
                 95% CI : (0.5757, 0.6825)
#Repetition 2 (set.seed(12345))
Prediction bad good
      bad   32   26
      good  67  205
```

```
               Accuracy : 0.7182
                 95% CI : (0.6663, 0.7661)
#Repetition 3 (set.seed(1234))
Prediction bad good
      bad   33   26
      good  66  205

               Accuracy : 0.7212
                 95% CI : (0.6695, 0.7689)
#Repetition 4 (set.seed(123))
Prediction bad good
      bad   25   38
      good  74  193

               Accuracy : 0.6606
                 95% CI : (0.6067, 0.7116)
```

The average error is coming out to be 0.32 with a sd of 0.04

2.1.2 Training error

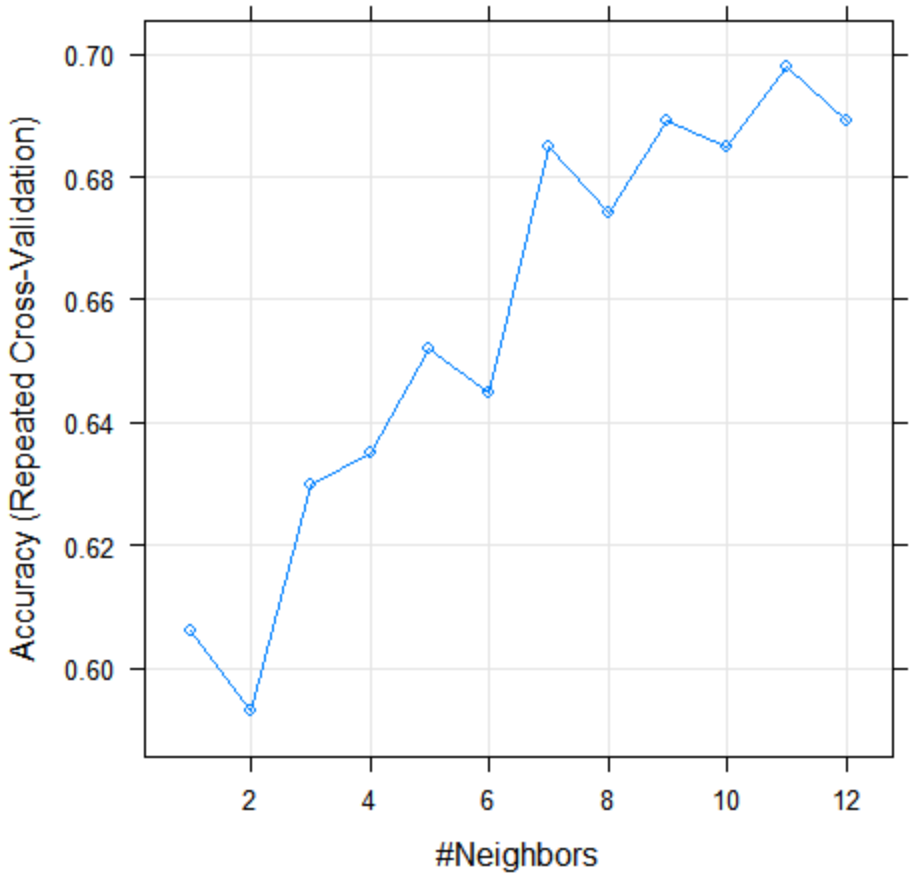
```
#Repetition 1 (set.seed(123456))
Prediction bad good
      bad   90   38
      good 111  431

               Accuracy : 0.7776
                 95% CI : (0.7442, 0.8086)

#Repetition 2 (set.seed(12345))
Prediction bad good
      bad   86   43
      good 115  426

               Accuracy : 0.7642
                 95% CI : (0.7302, 0.7958)

#Repetition 3 (set.seed(1234))
Prediction bad good
      bad   69   36
      good 132  433

               Accuracy : 0.7493
                 95% CI : (0.7146, 0.7817)
#Repetition 4 (set.seed(123))
Prediction bad good
      bad   91   41
      good 110  428

               Accuracy : 0.7746
                 95% CI : (0.7411, 0.8057)
```

The average error is 0.233 with a standard deviation of 0.01
Once again the training error is lower than the test error. However, the test error is a better estimate as the training error is overoptimistic

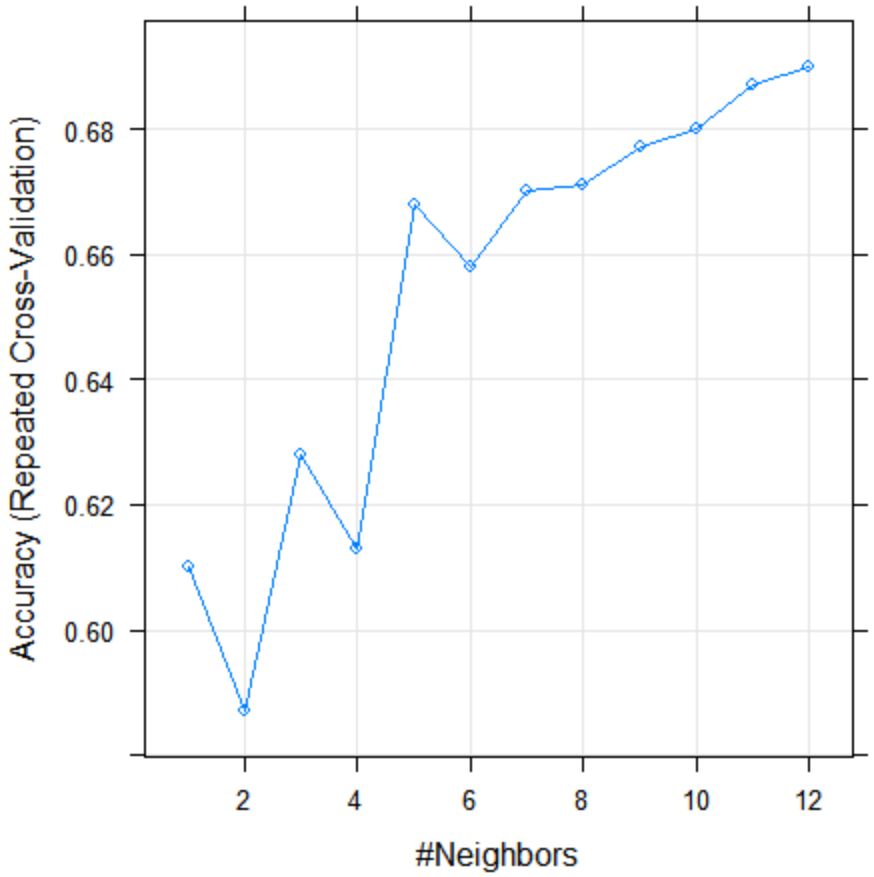
2.2 Cross Validatation

In order to tune our model the optimum number of k nearest neighbours was found.

2.2.1 Number of folds = 10

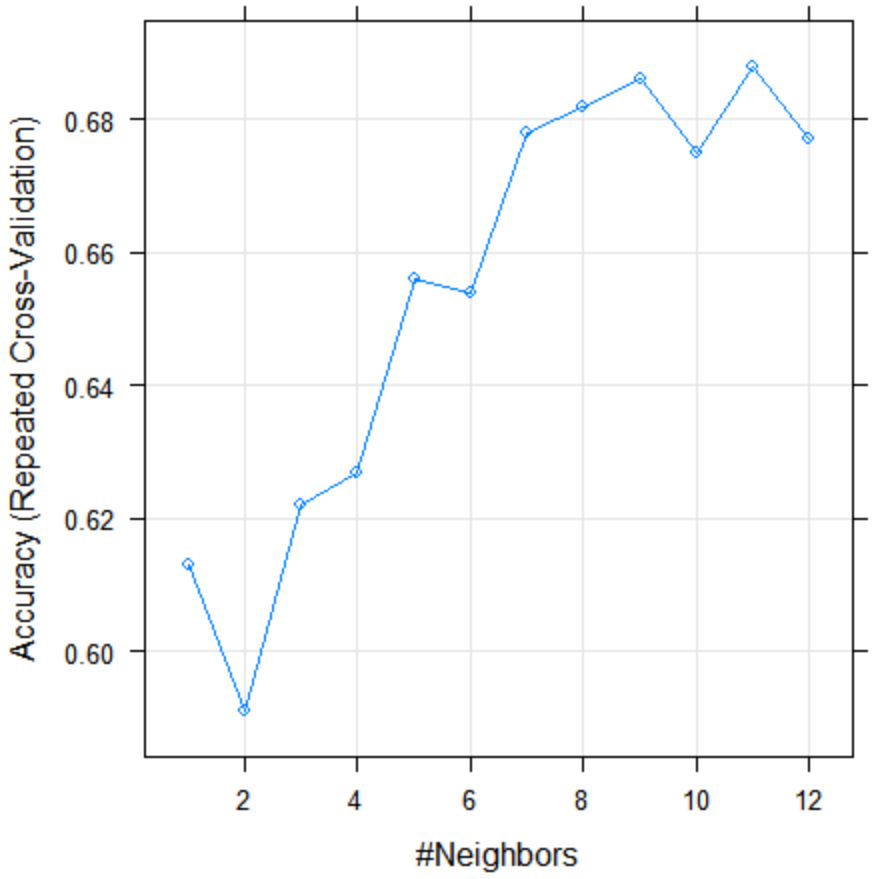



```
Best accuracy = 0.7, k = 11
2.2.2   Number of folds = 5
```



Best accuracy = 0.69, k = 12

2.2.3 Number of folds = 20




```
Best accuracy = 0.69, k = 11
```

The average error is 0.31 with a sd of 0.004. The CV error estimate gives the lowest variability as expected. In view of model simplicity and too avoid ove rgeneralizing, a k value of 7 is taken rather than k = 12. This gives average accuracy to be 0.68.

```
   3. Naïve Bayes

   3.1  Independent Test Set
   3.1.1       Test Error

#Repetition 1 (set.seed(123456))
Prediction bad good
      bad   50   36
      good  49  195

               Accuracy : 0.7424
                 95% CI : (0.6917, 0.7888)


#Repetition 2 (set.seed(12345))
Prediction bad good
      bad   42   41
      good  57  190
```

```
               Accuracy : 0.703
                 95% CI : (0.6505, 0.7518)

#Repetition 3 (set.seed(1234))
Prediction bad good
      bad   48   25
      good  51  206

               Accuracy : 0.7697
                 95% CI : (0.7204, 0.814)

#Repetition 4 (set.seed(123))
Prediction bad good
      bad   51   32
      good  48  199

               Accuracy : 0.7576
                 95% CI : (0.7076, 0.8028)
```

The average error is 0.26 and standard deviation is 0.025

3.1.2 Training error

```
#Repetition 1 (set.seed(123456))
Prediction bad good
      bad  108   59
      good  93  410

               Accuracy : 0.7731
                 95% CI : (0.7395, 0.8043)
#Repetition 2 (set.seed(12345))
Prediction bad good
      bad  113   56
      good  88  413

               Accuracy : 0.7851
                 95% CI : (0.752, 0.8156)
#Repetition 3 (set.seed(1234))
Prediction bad good
      bad  107   66
      good  94  403

               Accuracy : 0.7612
                 95% CI : (0.7271, 0.793)
#Repetition 4 (set.seed(123))
Prediction bad good
      bad  110   58
      good  91  411

               Accuracy : 0.7776
                 95% CI : (0.7442, 0.8086)
```

The average error is 0.23 and sd = 0.086

Thes test error is larger than the training error as expected

a. Cross Validatation

i. Number of folds = 10

```
usekernel  Accuracy   Kappa
  FALSE      0.6955556  0.3439128
   TRUE      0.7000000  0.0000000
```

ii. Number of folds = 5

```
usekernel  Accuracy  Kappa
  FALSE      0.685     0.3164143
   TRUE      0.700     0.0000000
```

iii. Number of folds = 20

```
usekernel  Accuracy  Kappa
    FALSE      0.68      0.3128611
     TRUE      0.70      0.0000000
```

For TRUE, there is no variability with an average error estimate of 0.3. For FALSE, the average error estimate is 0.31 with sd = 0.006 which is lower than independent test set.

Since the data size is small/moderate, CV estimate of error is taken along wi th fact that it gives the lowest variability. From the data in hand, it is ap parent that Decision Trees give the best result in terms of accuracy as expec ted.