FML Assignment 2

shiva gadila

2023-02-19

```
#Loading the packages

library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(ISLR)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(class)
library(FNN)
```

```
##
## Attaching package: 'FNN'
```

```
## The following objects are masked from 'package:class':
##
##     knn, knn.cv
```

```
# Importing the dataset.
UniversalBank <- read.csv("~/Desktop/UniversalBank.csv")
```

```
#Performing a K-NN classification with all attributes except ID and ZIP code.
UniversalBank$ID <- NULL
UniversalBank$ZIP.Code <- NULL
summary(UniversalBank)
```

```
##       Age           Experience       Income           Family
##  Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   :1.000
##  1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:1.000
##  Median :45.00   Median :20.0   Median : 64.00   Median :2.000
##  Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :2.396
##  3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:3.000
##  Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :4.000
##      CCAvg          Education        Mortgage     Personal.Loan
##  Min.   : 0.000   Min.   :1.000   Min.   :  0.0   Min.   :0.000
##  1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0   1st Qu.:0.000
##  Median : 1.500   Median :2.000   Median :  0.0   Median :0.000
##  Mean   : 1.938   Mean   :1.881   Mean   : 56.5   Mean   :0.096
##  3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0   3rd Qu.:0.000
##  Max.   :10.000   Max.   :3.000   Max.   :635.0   Max.   :1.000
##  Securities.Account   CD.Account         Online         CreditCard
##  Min.   :0.0000     Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000     Median :0.0000   Median :1.0000   Median :0.000
##  Mean   :0.1044     Mean   :0.0604   Mean   :0.5968   Mean   :0.294
##  3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.0000     Max.   :1.0000   Max.   :1.0000   Max.   :1.000
```

```
UniversalBank$Personal.Loan =  as.factor(UniversalBank$Personal.Loan)
```

```
#Creating dummy  variables
education_1 <- ifelse(UniversalBank$Education==1 ,1,0)

education_2 <- ifelse(UniversalBank$Education==2 ,1,0)

education_3 <- ifelse(UniversalBank$Education==3 ,1,0)

unibank<-data.frame(Age=UniversalBank$Age,Experience=UniversalBank$Experience,Income=UniversalBank$Income,Family=UniversalBank$Family,CCAvg=UniversalBank$CCAvg, education_1=education_1,education_2=education_2,education_3=education_3,Personal.Loan=UniversalBank$Personal.Loan,Mortgage=UniversalBank$Mortgage,Securities.Account=UniversalBank$Securities.Account,CD.Account=UniversalBank$CD.Account,Online=UniversalBank$Online,CreditCard=UniversalBank$CreditCard)
head(unibank)
```

```
##   Age Experience Income Family CCAvg education_1 education_2 education_3
## 1  25          1     49      4   1.6           1           0           0
## 2  45         19     34      3   1.5           1           0           0
## 3  39         15     11      1   1.0           1           0           0
## 4  35          9    100      1   2.7           0           1           0
## 5  35          8     45      4   1.0           0           1           0
## 6  37         13     29      4   0.4           0           1           0
##   Personal.Loan Mortgage Securities.Account CD.Account Online CreditCard
## 1             0        0                  1          0      0          0
## 2             0        0                  1          0      0          0
## 3             0        0                  0          0      0          0
## 4             0        0                  0          0      0          0
## 5             0        0                  0          0      0          1
## 6             0      155                  0          0      1          0
```

```
#Dividing into training and validation
Model.normalise <- preProcess(UniversalBank[, -8],method = c("center", "scale"))
summary(UniversalBank)
```

```
##       Age           Experience       Income           Family
##  Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   :1.000
##  1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:1.000
##  Median :45.00   Median :20.0   Median : 64.00   Median :2.000
##  Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :2.396
##  3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:3.000
##  Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :4.000
##      CCAvg           Education        Mortgage     Personal.Loan
##  Min.   : 0.000   Min.   :1.000   Min.   :  0.0   0:4520
##  1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0   1: 480
##  Median : 1.500   Median :2.000   Median :  0.0
##  Mean   : 1.938   Mean   :1.881   Mean   : 56.5
##  3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0
##  Max.   :10.000   Max.   :3.000   Max.   :635.0
##  Securities.Account   CD.Account         Online         CreditCard
##  Min.   :0.0000     Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000     Median :0.0000   Median :1.0000   Median :0.000
##  Mean   :0.1044     Mean   :0.0604   Mean   :0.5968   Mean   :0.294
##  3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.0000     Max.   :1.0000   Max.   :1.0000   Max.   :1.000
```

```
UniversalBank.normalise <- predict(Model.normalise,UniversalBank)
summary(UniversalBank.normalise)
```

```
##       Age             Experience            Income            Family
##  Min.   :-1.94871   Min.   :-2.014710   Min.   :-1.4288   Min.   :-1.2167
##  1st Qu.:-0.90188   1st Qu.:-0.881116   1st Qu.:-0.7554   1st Qu.:-1.2167
##  Median :-0.02952   Median :-0.009121   Median :-0.2123   Median :-0.3454
##  Mean   : 0.00000   Mean   : 0.000000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.84284   3rd Qu.: 0.862874   3rd Qu.: 0.5263   3rd Qu.: 0.5259
##  Max.   : 1.88967   Max.   : 1.996468   Max.   : 3.2634   Max.   : 1.3973
##      CCAvg           Education          Mortgage       Personal.Loan
##  Min.   :-1.1089   Min.   :-1.0490   Min.   :-0.5555   0:4520
##  1st Qu.:-0.7083   1st Qu.:-1.0490   1st Qu.:-0.5555   1: 480
##  Median :-0.2506   Median : 0.1417   Median :-0.5555
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.3216   3rd Qu.: 1.3324   3rd Qu.: 0.4375
##  Max.   : 4.6131   Max.   : 1.3324   Max.   : 5.6875
##  Securities.Account   CD.Account          Online          CreditCard
##  Min.   :-0.3414    Min.   :-0.2535   Min.   :-1.2165   Min.   :-0.6452
##  1st Qu.:-0.3414    1st Qu.:-0.2535   1st Qu.:-1.2165   1st Qu.:-0.6452
##  Median :-0.3414    Median :-0.2535   Median : 0.8219   Median :-0.6452
##  Mean   : 0.0000    Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.3414    3rd Qu.:-0.2535   3rd Qu.: 0.8219   3rd Qu.: 1.5495
##  Max.   : 2.9286    Max.   : 3.9438   Max.   : 0.8219   Max.   : 1.5495
```

```
Index_Train <- createDataPartition(UniversalBank$Personal.Loan, p = 0.6, list = FALSE)
Train = UniversalBank.normalise[Index_Train,]
validation = UniversalBank.normalise[-Index_Train,]
```

```
#QUESTION-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?
#Prediction of data
library(FNN)
to_Predict = data.frame(Age = 40, Experience = 10, Income = 84, Family = 2,
                        CCAvg = 2, Education = 1, Mortgage = 0, Securities.Account =
                          0, CD.Account = 0, Online = 1, CreditCard = 1)
print(to_Predict)
```

```
##   Age Experience Income Family CCAvg Education Mortgage Securities.Account
## 1  40         10     84      2     2         1        0                   0
##   CD.Account Online CreditCard
## 1          0      1          1
```

```
Predict.Normalise <- predict(Model.normalise,to_Predict)
Predictions <- knn(train= as.data.frame(Train[,1:7,9:12]),
                   test = as.data.frame(Predict.Normalise[,1:7,9:12]),
                   cl= Train$Personal.Loan,
                   k=1)
```

```
## Warning in drop && !has.j: 'length(x) = 4 > 1' in coercion to 'logical(1)'
```

```
## Warning in drop && length(y) == 1L: 'length(x) = 4 > 1' in coercion to
## 'logical(1)'
```

```
## Warning in drop && !mdrop: 'length(x) = 4 > 1' in coercion to 'logical(1)'
```

```
## Warning in drop && !has.j: 'length(x) = 4 > 1' in coercion to 'logical(1)'
```

```
## Warning in drop && length(y) == 1L: 'length(x) = 4 > 1' in coercion to
## 'logical(1)'
```

```
## Warning in drop && !mdrop: 'length(x) = 4 > 1' in coercion to 'logical(1)'
```

```
#QUESTION 2 - What is a choice of k that balances between overfitting and ignoring the predic
tor information?
set.seed(123)
UniversalBank <- trainControl(method= "repeatedcv", number = 3, repeats = 2)
searchGrid = expand.grid(k=1:10)
knn.model = train(Personal.Loan~., data = Train, method = 'knn', tuneGrid = searchGrid,trCont
rol = UniversalBank)
knn.model
```

```
## k-Nearest Neighbors
##
## 3000 samples
##   11 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (3 fold, repeated 2 times)
## Summary of sample sizes: 2000, 2000, 2000, 2000, 2000, 2000, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    1  0.9525000  0.6999607
##    2  0.9453333  0.6522815
##    3  0.9538333  0.6842781
##    4  0.9516667  0.6671257
##    5  0.9511667  0.6593635
##    6  0.9525000  0.6695893
##    7  0.9506667  0.6498425
##    8  0.9476667  0.6249816
##    9  0.9463333  0.6111616
##   10  0.9431667  0.5799230
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 3.
```

```
#The value of k is 3.This is the value that balances between overfitting and ignoring the pre
dictor information
```

```
#QUESTION 3- Show the confusion matrix for the validation data that results from using the be
st k.
UniversalBank_prediction <- predict(knn.model,validation)
confusionMatrix(UniversalBank_prediction,validation$Personal.Loan)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1796   61
##          1   12  131
##
##                Accuracy : 0.9635
##                  95% CI : (0.9543, 0.9713)
##     No Information Rate : 0.904
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7626
##
##  Mcnemar's Test P-Value : 1.932e-08
##
##             Sensitivity : 0.9934
##             Specificity : 0.6823
##          Pos Pred Value : 0.9672
##          Neg Pred Value : 0.9161
##              Prevalence : 0.9040
##          Detection Rate : 0.8980
##    Detection Prevalence : 0.9285
##       Balanced Accuracy : 0.8378
##
##        'Positive' Class : 0
##
```

```
#This matrix has a 95.9% accuracy.
#This the confusion matrix for the validation data that results from using the best k.
```

```
#QUESTION 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.

ForPredictNorm = data.frame(Age = 40, Experience = 10, Income = 84, Family = 2,
                            CCAvg = 2, Education = 1, Mortgage = 0,
                            Securities.Account =0, CD.Account = 0, Online = 1,
                            CreditCard = 1)
ForPredictNorm = predict(Model.normalise, ForPredictNorm)
predict(knn.model, ForPredictNorm)
```

```
## [1] 0
## Levels: 0 1
```

```
#It results in level 0,1
```

```
#QUESTION 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.

#Creating Training, Test, and validation sets from the data collection.

Train_size = 0.5 #training(50%)
Train_Index = createDataPartition(UniversalBank.normalise$Personal.Loan, p = 0.5, list = FALSE)
Train = UniversalBank.normalise[Train_Index,]

valid_size = 0.3 #validation(30%)
Validation_Index = createDataPartition(UniversalBank.normalise$Personal.Loan, p = 0.3, list = FALSE)
validation = UniversalBank.normalise[Validation_Index,]

Test_size = 0.2 #Test Data(20%)
Test_Index = createDataPartition(UniversalBank.normalise$Personal.Loan, p = 0.2, list = FALSE)
Test = UniversalBank.normalise[Test_Index,]


Trainingknn <- knn(train = Train[,-8], test = Train[,-8], cl = Train[,8], k =3)
Validknn <- knn(train = Train[,-8], test = validation[,-8], cl = Train[,8], k =3)
Testingknn <- knn(train = Train[,-8], test = Test[,-8], cl = Train[,8], k =3)

confusionMatrix(Trainingknn, Train[,8])
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2255   66
##          1    5  174
##
##                Accuracy : 0.9716
##                  95% CI : (0.9643, 0.9778)
##     No Information Rate : 0.904
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8154
##
##  Mcnemar's Test P-Value : 1.074e-12
##
##             Sensitivity : 0.9978
##             Specificity : 0.7250
##          Pos Pred Value : 0.9716
##          Neg Pred Value : 0.9721
##              Prevalence : 0.9040
##          Detection Rate : 0.9020
##    Detection Prevalence : 0.9284
##       Balanced Accuracy : 0.8614
##
##        'Positive' Class : 0
##
```

```
confusionMatrix(Validknn, validation[,8])
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1349   53
##          1    7   91
##
##                Accuracy : 0.96
##                  95% CI : (0.9488, 0.9693)
##     No Information Rate : 0.904
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7312
##
##  Mcnemar's Test P-Value : 6.267e-09
##
##             Sensitivity : 0.9948
##             Specificity : 0.6319
##          Pos Pred Value : 0.9622
##          Neg Pred Value : 0.9286
##              Prevalence : 0.9040
##          Detection Rate : 0.8993
##    Detection Prevalence : 0.9347
##       Balanced Accuracy : 0.8134
##
##        'Positive' Class : 0
##
```

```
confusionMatrix(Testingknn, Test[,8])
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 900  26
##          1   4  70
##
##                Accuracy : 0.97
##                  95% CI : (0.9574, 0.9797)
##     No Information Rate : 0.904
##     P-Value [Acc > NIR] : 3.048e-16
##
##                   Kappa : 0.8074
##
##  Mcnemar's Test P-Value : 0.000126
##
##             Sensitivity : 0.9956
##             Specificity : 0.7292
##          Pos Pred Value : 0.9719
##          Neg Pred Value : 0.9459
##              Prevalence : 0.9040
##          Detection Rate : 0.9000
##    Detection Prevalence : 0.9260
##       Balanced Accuracy : 0.8624
##
##        'Positive' Class : 0
##
```

```
# The accuracy for this knn model is 0.973 or 97.3%.
# The Sensitivity for this knn model is 0.9956 or 99.56%.
# The Specificity for this knn model is 0.7604 or 76.04%.
```