ML Assignment 2

Yash Bhanushali

2/20/2022

library(readr)  
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(fastDummies)

## Warning: package 'fastDummies' was built under R version 4.1.2

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(class)

library(readr)  
UniversalBank\_dataset <- read\_csv("D:/FML/UniversalBank (1).csv")

## Rows: 5000 Columns: 14

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## dbl (14): ID, Age, Experience, Income, ZIP Code, Family, CCAvg, Education, M...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

summary(UniversalBank\_dataset)

## ID Age Experience Income ZIP Code   
## Min. : 1 Min. :23.00 Min. :-3.0 Min. : 8.00 Min. : 9307   
## 1st Qu.:1251 1st Qu.:35.00 1st Qu.:10.0 1st Qu.: 39.00 1st Qu.:91911   
## Median :2500 Median :45.00 Median :20.0 Median : 64.00 Median :93437   
## Mean :2500 Mean :45.34 Mean :20.1 Mean : 73.77 Mean :93153   
## 3rd Qu.:3750 3rd Qu.:55.00 3rd Qu.:30.0 3rd Qu.: 98.00 3rd Qu.:94608   
## Max. :5000 Max. :67.00 Max. :43.0 Max. :224.00 Max. :96651   
## Family CCAvg Education Mortgage   
## Min. :1.000 Min. : 0.000 Min. :1.000 Min. : 0.0   
## 1st Qu.:1.000 1st Qu.: 0.700 1st Qu.:1.000 1st Qu.: 0.0   
## Median :2.000 Median : 1.500 Median :2.000 Median : 0.0   
## Mean :2.396 Mean : 1.938 Mean :1.881 Mean : 56.5   
## 3rd Qu.:3.000 3rd Qu.: 2.500 3rd Qu.:3.000 3rd Qu.:101.0   
## Max. :4.000 Max. :10.000 Max. :3.000 Max. :635.0   
## Personal Loan Securities Account CD Account Online   
## Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.000 Median :0.0000 Median :0.0000 Median :1.0000   
## Mean :0.096 Mean :0.1044 Mean :0.0604 Mean :0.5968   
## 3rd Qu.:0.000 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:1.0000   
## Max. :1.000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## CreditCard   
## Min. :0.000   
## 1st Qu.:0.000   
## Median :0.000   
## Mean :0.294   
## 3rd Qu.:1.000   
## Max. :1.000

##Datacleaning

#eliminating ID and ZipCode Columns from Dataset   
UniversalBank\_dataset <- UniversalBank\_dataset[,c(-1,-5)]  
str(UniversalBank\_dataset)

## tibble [5,000 x 12] (S3: tbl\_df/tbl/data.frame)  
## $ Age : num [1:5000] 25 45 39 35 35 37 53 50 35 34 ...  
## $ Experience : num [1:5000] 1 19 15 9 8 13 27 24 10 9 ...  
## $ Income : num [1:5000] 49 34 11 100 45 29 72 22 81 180 ...  
## $ Family : num [1:5000] 4 3 1 1 4 4 2 1 3 1 ...  
## $ CCAvg : num [1:5000] 1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...  
## $ Education : num [1:5000] 1 1 1 2 2 2 2 3 2 3 ...  
## $ Mortgage : num [1:5000] 0 0 0 0 0 155 0 0 104 0 ...  
## $ Personal Loan : num [1:5000] 0 0 0 0 0 0 0 0 0 1 ...  
## $ Securities Account: num [1:5000] 1 1 0 0 0 0 0 0 0 0 ...  
## $ CD Account : num [1:5000] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Online : num [1:5000] 0 0 0 0 0 1 1 0 1 0 ...  
## $ CreditCard : num [1:5000] 0 0 0 0 1 0 0 1 0 0 ...

#As personal loan is predictive variable so converting it to factor  
UniversalBank\_dataset$`Personal Loan` <-as.factor(UniversalBank\_dataset$`Personal Loan`)  
UniversalBank\_dataset$Education <-as.factor(UniversalBank\_dataset$Education)  
View(UniversalBank\_dataset)  
#Dummying  
library(fastDummies)  
UniversalBank\_dataset\_d <- dummy\_cols(UniversalBank\_dataset %>% select(-`Personal Loan`))  
UniversalBank\_dataset\_d <- UniversalBank\_dataset\_d %>% select(-Education) %>%   
 mutate(`Personal Loan` = UniversalBank\_dataset$`Personal Loan`)

##Data Partition and preprocessing

set.seed(300)  
index <- createDataPartition(UniversalBank\_dataset\_d$`Personal Loan`, p=0.5, list = FALSE)  
UniversalBank\_dataset\_train\_df <- UniversalBank\_dataset\_d[index,]  
UniversalBank\_dataset\_test\_df <- UniversalBank\_dataset\_d[-index,]  
#normalize the data.  
scale\_fun <- preProcess(UniversalBank\_dataset\_train\_df[,-13], method = c("center", "scale"))  
UniversalBank\_dataset\_train\_norm <- predict(scale\_fun, UniversalBank\_dataset\_train\_df[,-13])  
UniversalBank\_dataset\_test\_norm <- predict(scale\_fun, UniversalBank\_dataset\_test\_df[,-13])  
dim(UniversalBank\_dataset\_train\_norm)

## [1] 2500 13

summary(UniversalBank\_dataset\_train\_norm)

## Age Experience Income Family   
## Min. :-1.916125 Min. :-1.98017 Min. :-1.4313 Min. :-1.2325   
## 1st Qu.:-0.872977 1st Qu.:-0.85114 1st Qu.:-0.7516 1st Qu.:-1.2325   
## Median :-0.003686 Median : 0.01733 Median :-0.2035 Median :-0.3614   
## Mean : 0.000000 Mean : 0.00000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.865605 3rd Qu.: 0.88581 3rd Qu.: 0.4762 3rd Qu.: 0.5098   
## Max. : 1.908754 Max. : 2.01484 Max. : 3.1730 Max. : 1.3810   
## CCAvg Mortgage Securities Account CD Account   
## Min. :-1.1068 Min. :-0.5691 Min. :-0.3421 Min. :-0.2517   
## 1st Qu.:-0.7015 1st Qu.:-0.5691 1st Qu.:-0.3421 1st Qu.:-0.2517   
## Median :-0.2383 Median :-0.5691 Median :-0.3421 Median :-0.2517   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.3407 3rd Qu.: 0.4425 3rd Qu.:-0.3421 3rd Qu.:-0.2517   
## Max. : 4.2782 Max. : 5.4758 Max. : 2.9221 Max. : 3.9714   
## Online CreditCard Education\_1 Education\_2   
## Min. :-1.227 Min. :-0.6439 Min. :-0.8508 Min. :-0.6235   
## 1st Qu.:-1.227 1st Qu.:-0.6439 1st Qu.:-0.8508 1st Qu.:-0.6235   
## Median : 0.815 Median :-0.6439 Median :-0.8508 Median :-0.6235   
## Mean : 0.000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.815 3rd Qu.: 1.5523 3rd Qu.: 1.1749 3rd Qu.: 1.6032   
## Max. : 0.815 Max. : 1.5523 Max. : 1.1749 Max. : 1.6032   
## Personal Loan  
## 0:2260   
## 1: 240   
##   
##   
##   
##

summary(UniversalBank\_dataset\_test\_norm)

## Age Experience Income Family   
## Min. :-1.91613 Min. :-1.98017 Min. :-1.4313 Min. :-1.23251   
## 1st Qu.:-0.78605 1st Qu.:-0.76430 1st Qu.:-0.7516 1st Qu.:-1.23251   
## Median : 0.08324 Median : 0.10418 Median :-0.2035 Median :-0.36136   
## Mean : 0.05146 Mean : 0.05284 Mean : 0.0216 Mean :-0.03206   
## 3rd Qu.: 0.86560 3rd Qu.: 0.88581 3rd Qu.: 0.5694 3rd Qu.: 0.50980   
## Max. : 1.90875 Max. : 2.01484 Max. : 3.3045 Max. : 1.38096   
## CCAvg Mortgage Securities Account CD Account   
## Min. :-1.10684 Min. :-0.56911 Min. :-0.342085 Min. :-0.251698   
## 1st Qu.:-0.70152 1st Qu.:-0.56911 1st Qu.:-0.342085 1st Qu.:-0.251698   
## Median :-0.18038 Median :-0.56911 Median :-0.342085 Median :-0.251698   
## Mean : 0.03059 Mean :-0.03115 Mean :-0.002611 Mean : 0.006757   
## 3rd Qu.: 0.39865 3rd Qu.: 0.39102 3rd Qu.:-0.342085 3rd Qu.:-0.251698   
## Max. : 4.68350 Max. : 5.65217 Max. : 2.922083 Max. : 3.971424   
## Online CreditCard Education\_1 Education\_2   
## Min. :-1.22654 Min. :-0.643942 Min. :-0.850793 Min. :-0.623485   
## 1st Qu.:-1.22654 1st Qu.:-0.643942 1st Qu.:-0.850793 1st Qu.:-0.623485   
## Median : 0.81497 Median :-0.643942 Median :-0.850793 Median :-0.623485   
## Mean :-0.01633 Mean : 0.003514 Mean :-0.003241 Mean : 0.002672   
## 3rd Qu.: 0.81497 3rd Qu.: 1.552313 3rd Qu.: 1.174904 3rd Qu.: 1.603247   
## Max. : 0.81497 Max. : 1.552313 Max. : 1.174904 Max. : 1.603247   
## Personal Loan  
## 0:2260   
## 1: 240   
##   
##   
##   
##

##KNN Modeling #1. Predicting the Customer with K=1

#Predicting the Customer with K=1  
Q1 <- data.frame(40, 10, 84, 2, 2, 0, 1, 0, 0, 0, 0, 1, 1)  
knn\_prediction <- knn(UniversalBank\_dataset\_train\_norm, Q1, cl=UniversalBank\_dataset\_train\_df$`Personal Loan`, k=1, prob = 0.6)  
knn\_prediction

## [1] 1  
## attr(,"prob")  
## [1] 1  
## Levels: 0 1

#2. Choosing value of k

accuracy.df <- data.frame(k = seq(1, 13, 1), accuracy = rep(0, 13))  
for(i in 1:13) {  
 knn <- knn(UniversalBank\_dataset\_train\_norm, UniversalBank\_dataset\_test\_norm, cl = UniversalBank\_dataset\_train\_df$`Personal Loan`, k = i)  
 accuracy.df[i, 2] <- confusionMatrix(knn, UniversalBank\_dataset\_test\_df$`Personal Loan`)$overall[1]   
 }  
accuracy.df

## k accuracy  
## 1 1 0.9776  
## 2 2 0.9720  
## 3 3 0.9700  
## 4 4 0.9668  
## 5 5 0.9672  
## 6 6 0.9668  
## 7 7 0.9632  
## 8 8 0.9632  
## 9 9 0.9600  
## 10 10 0.9576  
## 11 11 0.9572  
## 12 12 0.9572  
## 13 13 0.9556

which.max( (accuracy.df$accuracy) ) #Here, our optimal k is 3

## [1] 1

#3. Validating data using the best ‘k’.

knn.pred3 <- knn(UniversalBank\_dataset\_train\_norm,UniversalBank\_dataset\_test\_norm,cl=UniversalBank\_dataset\_train\_df$`Personal Loan`,k=4,prob = TRUE)  
confusionMatrix(knn.pred3,UniversalBank\_dataset\_test\_df$`Personal Loan`)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 2255 73  
## 1 5 167  
##   
## Accuracy : 0.9688   
## 95% CI : (0.9612, 0.9753)  
## No Information Rate : 0.904   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.7942   
##   
## Mcnemar's Test P-Value : 3.293e-14   
##   
## Sensitivity : 0.9978   
## Specificity : 0.6958   
## Pos Pred Value : 0.9686   
## Neg Pred Value : 0.9709   
## Prevalence : 0.9040   
## Detection Rate : 0.9020   
## Detection Prevalence : 0.9312   
## Balanced Accuracy : 0.8468   
##   
## 'Positive' Class : 0   
##

#4. Classifying the customer using the best ‘k’

knn.pred4 <- knn(UniversalBank\_dataset\_train\_norm, Q1, cl=UniversalBank\_dataset\_train\_df$`Personal Loan`, k=4, prob = TRUE)  
knn.pred4

## [1] 1  
## attr(,"prob")  
## [1] 1  
## Levels: 0 1

#5. Repartitioning the data into training, validation, and test sets (50% : 30% : 20%)

set.seed(400)  
index\_b <- createDataPartition(UniversalBank\_dataset\_d$`Personal Loan`, p=0.5, list = FALSE)  
UniversalBank\_dataset\_training\_df2 <- UniversalBank\_dataset\_d[index\_b,]  
validation\_test\_idx <- UniversalBank\_dataset\_d[-index\_b,]  
validation\_test\_idx\_b <- createDataPartition(validation\_test\_idx$`Personal Loan`, p=0.6, list = FALSE)  
UniversalBank\_dataset\_val\_df2 <- validation\_test\_idx[validation\_test\_idx\_b,]  
UniversalBank\_dataset\_test\_df2 <- validation\_test\_idx[-validation\_test\_idx\_b,]  
#normalizing the data.  
scl\_fun\_b <- preProcess(UniversalBank\_dataset\_training\_df2[,-13], method = c("center", "scale"))  
UniversalBank\_dataset\_training\_norm2 <- predict(scl\_fun\_b, UniversalBank\_dataset\_training\_df2[,-13])  
UniversalBank\_dataset\_val\_norm2 <- predict(scl\_fun\_b, UniversalBank\_dataset\_val\_df2[,-13])  
UniversalBank\_dataset\_test\_norm2 <- predict(scl\_fun\_b, UniversalBank\_dataset\_test\_df2[,-13])  
knn.pred5 <- knn(UniversalBank\_dataset\_training\_norm2, UniversalBank\_dataset\_val\_norm2 , cl=UniversalBank\_dataset\_training\_df2$`Personal Loan`, k=4, prob = TRUE)  
confusionMatrix(knn.pred5,UniversalBank\_dataset\_val\_df2$`Personal Loan`)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 1350 42  
## 1 6 102  
##   
## Accuracy : 0.968   
## 95% CI : (0.9578, 0.9763)  
## No Information Rate : 0.904   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.7924   
##   
## Mcnemar's Test P-Value : 4.376e-07   
##   
## Sensitivity : 0.9956   
## Specificity : 0.7083   
## Pos Pred Value : 0.9698   
## Neg Pred Value : 0.9444   
## Prevalence : 0.9040   
## Detection Rate : 0.9000   
## Detection Prevalence : 0.9280   
## Balanced Accuracy : 0.8520   
##   
## 'Positive' Class : 0   
##