Assignment

Sai Prasad

2024-02-18

#Importing libraries of class, dpylr, caret & Dataset. Created a dataset called as “dataset\_universal”

library(class)  
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(caret)

## Loading required package: ggplot2

## Loading required package: lattice

dataset\_universal <- read.csv("C:\\Users\\desineni\\OneDrive - Kent State University\\Desktop\\boyapati\\sem\_1\\fml\\UniversalBank.csv")

Excluding unnecessory columns from the above dataset i.e,, ID, ZIP.Code

dataset\_universal1 <- dataset\_universal[,-1]  
dataset\_universal1 <- dataset\_universal1[,-4]

Conveting Personal.Loan into a factor

dataset\_universal$Personal.Loan=as.factor(dataset\_universal$Personal.Loan)  
head(dataset\_universal$Personal.Loan)

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

Executing above dataset with .na if there is any NA values

head(is.na(dataset\_universal1))

## Age Experience Income Family CCAvg Education Mortgage Personal.Loan  
## [1,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [2,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [3,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [4,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [5,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [6,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## Securities.Account CD.Account Online CreditCard  
## [1,] FALSE FALSE FALSE FALSE  
## [2,] FALSE FALSE FALSE FALSE  
## [3,] FALSE FALSE FALSE FALSE  
## [4,] FALSE FALSE FALSE FALSE  
## [5,] FALSE FALSE FALSE FALSE  
## [6,] FALSE FALSE FALSE FALSE

any(is.na(dataset\_universal1))

## [1] FALSE

Converting categorical variable into i.e education into dummy variables

Here we are converting Education into factor. There are three levels in Education

education = as.character(dataset\_universal1$Education)  
  
head(dataset\_universal1)

## Age Experience Income Family CCAvg Education Mortgage Personal.Loan  
## 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 2 0 0  
## 5 35 8 45 4 1.0 2 0 0  
## 6 37 13 29 4 0.4 2 155 0  
## Securities.Account CD.Account Online CreditCard  
## 1 1 0 0 0  
## 2 1 0 0 0  
## 3 0 0 0 0  
## 4 0 0 0 0  
## 5 0 0 0 1  
## 6 0 0 1 0

dataset\_universal2 <- cbind(dataset\_universal1[,-6],education)  
head(dataset\_universal2)

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

dummymodel<-dummyVars("~education",data = dataset\_universal2)  
educationdummy<-data.frame(predict(dummymodel,dataset\_universal2))  
head(educationdummy)

## education1 education2 education3  
## 1 1 0 0  
## 2 1 0 0  
## 3 1 0 0  
## 4 0 1 0  
## 5 0 1 0  
## 6 0 1 0

dataset\_universal\_dummy<-cbind(dataset\_universal2[,-12],educationdummy)  
head(dataset\_universal\_dummy)

## Age Experience Income Family CCAvg Mortgage Personal.Loan Securities.Account  
## 1 25 1 49 4 1.6 0 0 1  
## 2 45 19 34 3 1.5 0 0 1  
## 3 39 15 11 1 1.0 0 0 0  
## 4 35 9 100 1 2.7 0 0 0  
## 5 35 8 45 4 1.0 0 0 0  
## 6 37 13 29 4 0.4 155 0 0  
## CD.Account Online CreditCard education1 education2 education3  
## 1 0 0 0 1 0 0  
## 2 0 0 0 1 0 0  
## 3 0 0 0 1 0 0  
## 4 0 0 0 0 1 0  
## 5 0 0 1 0 1 0  
## 6 0 1 0 0 1 0

set.seed(123)  
train<-createDataPartition(dataset\_universal\_dummy$Personal.Loan,p=0.60,list = FALSE)  
train\_set<-dataset\_universal\_dummy[train,]  
nrow(train\_set)

## [1] 3000

validation\_set<-dataset\_universal\_dummy[-train,]  
nrow(validation\_set)

## [1] 2000

test\_set<-data.frame(Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Mortgage = 0, Securities.Account = 0, CD.Account = 0, Online = 1,   
 CreditCard = 1,education1 = 0, education2 = 1, education3 = 0)  
  
  
summary(train\_set)

## Age Experience Income Family   
## Min. :23.00 Min. :-3.00 Min. : 8.0 Min. :1.000   
## 1st Qu.:36.00 1st Qu.:10.00 1st Qu.: 39.0 1st Qu.:1.000   
## Median :46.00 Median :20.00 Median : 64.0 Median :2.000   
## Mean :45.43 Mean :20.21 Mean : 73.4 Mean :2.399   
## 3rd Qu.:55.00 3rd Qu.:30.00 3rd Qu.: 95.0 3rd Qu.:4.000   
## Max. :67.00 Max. :43.00 Max. :224.0 Max. :4.000   
## CCAvg Mortgage Personal.Loan Securities.Account  
## Min. : 0.000 Min. : 0.00 Min. :0.00000 Min. :0.000   
## 1st Qu.: 0.700 1st Qu.: 0.00 1st Qu.:0.00000 1st Qu.:0.000   
## Median : 1.500 Median : 0.00 Median :0.00000 Median :0.000   
## Mean : 1.924 Mean : 55.29 Mean :0.09267 Mean :0.103   
## 3rd Qu.: 2.600 3rd Qu.: 98.00 3rd Qu.:0.00000 3rd Qu.:0.000   
## Max. :10.000 Max. :635.00 Max. :1.00000 Max. :1.000   
## CD.Account Online CreditCard education1   
## Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000   
## Median :0.000 Median :1.0000 Median :0.0000 Median :0.000   
## Mean :0.059 Mean :0.5997 Mean :0.2943 Mean :0.424   
## 3rd Qu.:0.000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :1.000 Max. :1.0000 Max. :1.0000 Max. :1.000   
## education2 education3   
## Min. :0.000 Min. :0.000   
## 1st Qu.:0.000 1st Qu.:0.000   
## Median :0.000 Median :0.000   
## Mean :0.285 Mean :0.291   
## 3rd Qu.:1.000 3rd Qu.:1.000   
## Max. :1.000 Max. :1.000

summary(validation\_set)

## Age Experience Income Family   
## Min. :23.0 Min. :-3.00 Min. : 8.00 Min. :1.000   
## 1st Qu.:35.0 1st Qu.:10.00 1st Qu.: 39.00 1st Qu.:1.000   
## Median :45.0 Median :20.00 Median : 63.00 Median :2.000   
## Mean :45.2 Mean :19.95 Mean : 74.34 Mean :2.393   
## 3rd Qu.:55.0 3rd Qu.:30.00 3rd Qu.:100.00 3rd Qu.:3.000   
## Max. :67.0 Max. :43.00 Max. :204.00 Max. :4.000   
## CCAvg Mortgage Personal.Loan Securities.Account  
## Min. : 0.000 Min. : 0.00 Min. :0.000 Min. :0.0000   
## 1st Qu.: 0.700 1st Qu.: 0.00 1st Qu.:0.000 1st Qu.:0.0000   
## Median : 1.600 Median : 0.00 Median :0.000 Median :0.0000   
## Mean : 1.959 Mean : 58.32 Mean :0.101 Mean :0.1065   
## 3rd Qu.: 2.500 3rd Qu.:103.00 3rd Qu.:0.000 3rd Qu.:0.0000   
## Max. :10.000 Max. :601.00 Max. :1.000 Max. :1.0000   
## CD.Account Online CreditCard education1   
## 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 :1.0000 Median :0.0000 Median :0.000   
## Mean :0.0625 Mean :0.5925 Mean :0.2935 Mean :0.412   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.000   
## education2 education3   
## Min. :0.000 Min. :0.000   
## 1st Qu.:0.000 1st Qu.:0.000   
## Median :0.000 Median :0.000   
## Mean :0.274 Mean :0.314   
## 3rd Qu.:1.000 3rd Qu.:1.000   
## Max. :1.000 Max. :1.000

summary(test\_set)

## Age Experience Income Family CCAvg Mortgage  
## Min. :40 Min. :10 Min. :84 Min. :2 Min. :2 Min. :0   
## 1st Qu.:40 1st Qu.:10 1st Qu.:84 1st Qu.:2 1st Qu.:2 1st Qu.:0   
## Median :40 Median :10 Median :84 Median :2 Median :2 Median :0   
## Mean :40 Mean :10 Mean :84 Mean :2 Mean :2 Mean :0   
## 3rd Qu.:40 3rd Qu.:10 3rd Qu.:84 3rd Qu.:2 3rd Qu.:2 3rd Qu.:0   
## Max. :40 Max. :10 Max. :84 Max. :2 Max. :2 Max. :0   
## Securities.Account CD.Account Online CreditCard education1  
## Min. :0 Min. :0 Min. :1 Min. :1 Min. :0   
## 1st Qu.:0 1st Qu.:0 1st Qu.:1 1st Qu.:1 1st Qu.:0   
## Median :0 Median :0 Median :1 Median :1 Median :0   
## Mean :0 Mean :0 Mean :1 Mean :1 Mean :0   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:1 3rd Qu.:1 3rd Qu.:0   
## Max. :0 Max. :0 Max. :1 Max. :1 Max. :0   
## education2 education3  
## Min. :1 Min. :0   
## 1st Qu.:1 1st Qu.:0   
## Median :1 Median :0   
## Mean :1 Mean :0   
## 3rd Qu.:1 3rd Qu.:0   
## Max. :1 Max. :0

normvar<-c('Age',"Experience","Income","Family","CCAvg","Mortgage","Securities.Account","CD.Account","Online","CreditCard","education1","education2","education3")  
normalization\_values<-preProcess(train\_set[,normvar],method = c('center','scale'))  
  
trainset\_norm<-predict(normalization\_values,train\_set)  
summary(trainset\_norm)

## Age Experience Income Family   
## Min. :-1.96925 Min. :-2.03259 Min. :-1.4286 Min. :-1.2142   
## 1st Qu.:-0.82784 1st Qu.:-0.89392 1st Qu.:-0.7514 1st Qu.:-1.2142   
## Median : 0.05016 Median :-0.01801 Median :-0.2053 Median :-0.3463   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.84037 3rd Qu.: 0.85789 3rd Qu.: 0.4720 3rd Qu.: 1.3895   
## Max. : 1.89398 Max. : 1.99656 Max. : 3.2901 Max. : 1.3895   
## CCAvg Mortgage Personal.Loan Securities.Account  
## Min. :-1.1157 Min. :-0.5477 Min. :0.00000 Min. :-0.3388   
## 1st Qu.:-0.7098 1st Qu.:-0.5477 1st Qu.:0.00000 1st Qu.:-0.3388   
## Median :-0.2459 Median :-0.5477 Median :0.00000 Median :-0.3388   
## Mean : 0.0000 Mean : 0.0000 Mean :0.09267 Mean : 0.0000   
## 3rd Qu.: 0.3921 3rd Qu.: 0.4231 3rd Qu.:0.00000 3rd Qu.:-0.3388   
## Max. : 4.6835 Max. : 5.7429 Max. :1.00000 Max. : 2.9506   
## CD.Account Online CreditCard education1   
## Min. :-0.2504 Min. :-1.2237 Min. :-0.6457 Min. :-0.8578   
## 1st Qu.:-0.2504 1st Qu.:-1.2237 1st Qu.:-0.6457 1st Qu.:-0.8578   
## Median :-0.2504 Median : 0.8169 Median :-0.6457 Median :-0.8578   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.2504 3rd Qu.: 0.8169 3rd Qu.: 1.5481 3rd Qu.: 1.1653   
## Max. : 3.9930 Max. : 0.8169 Max. : 1.5481 Max. : 1.1653   
## education2 education3   
## Min. :-0.6312 Min. :-0.6405   
## 1st Qu.:-0.6312 1st Qu.:-0.6405   
## Median :-0.6312 Median :-0.6405   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 1.5836 3rd Qu.: 1.5606   
## Max. : 1.5836 Max. : 1.5606

validationset\_norm<-predict(normalization\_values,validation\_set)  
summary(validationset\_norm)

## Age Experience Income Family   
## Min. :-1.96925 Min. :-2.03259 Min. :-1.42861 Min. :-1.214194   
## 1st Qu.:-0.91565 1st Qu.:-0.89392 1st Qu.:-0.75140 1st Qu.:-1.214194   
## Median :-0.03764 Median :-0.01801 Median :-0.22710 Median :-0.346293   
## Mean :-0.01981 Mean :-0.02213 Mean : 0.02067 Mean :-0.005641   
## 3rd Qu.: 0.84037 3rd Qu.: 0.85789 3rd Qu.: 0.58119 3rd Qu.: 0.521609   
## Max. : 1.89398 Max. : 1.99656 Max. : 2.85314 Max. : 1.389511   
## CCAvg Mortgage Personal.Loan Securities.Account  
## Min. :-1.11573 Min. :-0.54769 Min. :0.000 Min. :-0.33881   
## 1st Qu.:-0.70979 1st Qu.:-0.54769 1st Qu.:0.000 1st Qu.:-0.33881   
## Median :-0.18786 Median :-0.54769 Median :0.000 Median :-0.33881   
## Mean : 0.02029 Mean : 0.03003 Mean :0.101 Mean : 0.01151   
## 3rd Qu.: 0.33407 3rd Qu.: 0.47267 3rd Qu.:0.000 3rd Qu.:-0.33881   
## Max. : 4.68347 Max. : 5.40606 Max. :1.000 Max. : 2.95057   
## CD.Account Online CreditCard education1   
## Min. :-0.25036 Min. :-1.22369 Min. :-0.645725 Min. :-0.85783   
## 1st Qu.:-0.25036 1st Qu.:-1.22369 1st Qu.:-0.645725 1st Qu.:-0.85783   
## Median :-0.25036 Median : 0.81693 Median :-0.645725 Median :-0.85783   
## Mean : 0.01485 Mean :-0.01462 Mean :-0.001828 Mean :-0.02428   
## 3rd Qu.:-0.25036 3rd Qu.: 0.81693 3rd Qu.: 1.548131 3rd Qu.: 1.16535   
## Max. : 3.99297 Max. : 0.81693 Max. : 1.548131 Max. : 1.16535   
## education2 education3   
## Min. :-0.63124 Min. :-0.64055   
## 1st Qu.:-0.63124 1st Qu.:-0.64055   
## Median :-0.63124 Median :-0.64055   
## Mean :-0.02436 Mean : 0.05063   
## 3rd Qu.: 1.58365 3rd Qu.: 1.56065   
## Max. : 1.58365 Max. : 1.56065

testset\_norm<-predict(normalization\_values,test\_set)  
summary(testset\_norm)

## Age Experience Income Family   
## Min. :-0.4766 Min. :-0.8939 Min. :0.2317 Min. :-0.3463   
## 1st Qu.:-0.4766 1st Qu.:-0.8939 1st Qu.:0.2317 1st Qu.:-0.3463   
## Median :-0.4766 Median :-0.8939 Median :0.2317 Median :-0.3463   
## Mean :-0.4766 Mean :-0.8939 Mean :0.2317 Mean :-0.3463   
## 3rd Qu.:-0.4766 3rd Qu.:-0.8939 3rd Qu.:0.2317 3rd Qu.:-0.3463   
## Max. :-0.4766 Max. :-0.8939 Max. :0.2317 Max. :-0.3463   
## CCAvg Mortgage Securities.Account CD.Account   
## Min. :0.04411 Min. :-0.5477 Min. :-0.3388 Min. :-0.2504   
## 1st Qu.:0.04411 1st Qu.:-0.5477 1st Qu.:-0.3388 1st Qu.:-0.2504   
## Median :0.04411 Median :-0.5477 Median :-0.3388 Median :-0.2504   
## Mean :0.04411 Mean :-0.5477 Mean :-0.3388 Mean :-0.2504   
## 3rd Qu.:0.04411 3rd Qu.:-0.5477 3rd Qu.:-0.3388 3rd Qu.:-0.2504   
## Max. :0.04411 Max. :-0.5477 Max. :-0.3388 Max. :-0.2504   
## Online CreditCard education1 education2   
## Min. :0.8169 Min. :1.548 Min. :-0.8578 Min. :1.584   
## 1st Qu.:0.8169 1st Qu.:1.548 1st Qu.:-0.8578 1st Qu.:1.584   
## Median :0.8169 Median :1.548 Median :-0.8578 Median :1.584   
## Mean :0.8169 Mean :1.548 Mean :-0.8578 Mean :1.584   
## 3rd Qu.:0.8169 3rd Qu.:1.548 3rd Qu.:-0.8578 3rd Qu.:1.584   
## Max. :0.8169 Max. :1.548 Max. :-0.8578 Max. :1.584   
## education3   
## Min. :-0.6405   
## 1st Qu.:-0.6405   
## Median :-0.6405   
## Mean :-0.6405   
## 3rd Qu.:-0.6405   
## Max. :-0.6405

#1: Classifying the given customer data

set.seed(123)  
newgrid<-expand.grid(k=c(1))  
newmodel<-train(Personal.Loan~.,data=trainset\_norm,method="knn",tuneGrid=newgrid)

## Warning in train.default(x, y, weights = w, ...): You are trying to do  
## regression and your outcome only has two possible values Are you trying to do  
## classification? If so, use a 2 level factor as your outcome column.

newmodel

## k-Nearest Neighbors   
##   
## 3000 samples  
## 13 predictor  
##   
## No pre-processing  
## Resampling: Bootstrapped (25 reps)   
## Summary of sample sizes: 3000, 3000, 3000, 3000, 3000, 3000, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.2143273 0.5011429 0.04604842  
##   
## Tuning parameter 'k' was held constant at a value of 1

predicting\_test<-predict(newmodel,testset\_norm)  
predicting\_test

## [1] 0

#Question 2: What is a choice of k that balances between overfiƫng and ignoring the predictor information? #identifying the best k and finding the best k

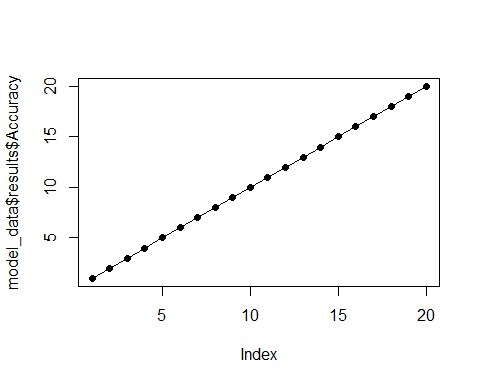
set.seed(123)  
searchGrid <- expand.grid(k=seq(1:20))  
model\_data<-train(Personal.Loan~.,data=trainset\_norm,method="knn",tuneGrid=searchGrid)

## Warning in train.default(x, y, weights = w, ...): You are trying to do  
## regression and your outcome only has two possible values Are you trying to do  
## classification? If so, use a 2 level factor as your outcome column.

model\_data

## k-Nearest Neighbors   
##   
## 3000 samples  
## 13 predictor  
##   
## No pre-processing  
## Resampling: Bootstrapped (25 reps)   
## Summary of sample sizes: 3000, 3000, 3000, 3000, 3000, 3000, ...   
## Resampling results across tuning parameters:  
##   
## k RMSE Rsquared MAE   
## 1 0.2143273 0.5011429 0.04604842  
## 2 0.2090459 0.5109643 0.05036350  
## 3 0.2027568 0.5313138 0.05279874  
## 4 0.1991433 0.5451486 0.05472585  
## 5 0.1962845 0.5585983 0.05612246  
## 6 0.1954729 0.5637810 0.05806894  
## 7 0.1956585 0.5651541 0.06019111  
## 8 0.1961477 0.5649393 0.06182045  
## 9 0.1962560 0.5668889 0.06326022  
## 10 0.1972181 0.5656002 0.06482543  
## 11 0.1979429 0.5651521 0.06615551  
## 12 0.1988287 0.5641083 0.06739236  
## 13 0.1997688 0.5612810 0.06853131  
## 14 0.2007159 0.5594330 0.06934727  
## 15 0.2014621 0.5582971 0.07037266  
## 16 0.2020913 0.5576605 0.07112699  
## 17 0.2028419 0.5562359 0.07181930  
## 18 0.2031976 0.5572771 0.07236332  
## 19 0.2036927 0.5570612 0.07288449  
## 20 0.2041528 0.5569869 0.07346275  
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final value used for the model was k = 6.

plot(model\_data$results$k,model\_data$results$Accuracy, type = 'o',pch = 16)



best\_k <- model\_data$bestTune[[1]]  
best\_k

## [1] 6

#Question 3:Confusion matrix for the validation data that results from using the best k

library(gmodels)  
  
train\_label<-trainset\_norm[,7]  
validation\_label<-validationset\_norm[,7]  
test\_label<-testset\_norm[,7]  
  
predicted\_validationlabel<-knn(trainset\_norm,validationset\_norm,cl=train\_label,k=5)  
  
CrossTable(x=validation\_label,y=predicted\_validationlabel,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 2000   
##   
##   
## | predicted\_validationlabel   
## validation\_label | 0 | 1 | Row Total |   
## -----------------|-----------|-----------|-----------|  
## 0 | 1795 | 3 | 1798 |   
## | 0.998 | 0.002 | 0.899 |   
## | 0.966 | 0.021 | |   
## | 0.897 | 0.002 | |   
## -----------------|-----------|-----------|-----------|  
## 1 | 63 | 139 | 202 |   
## | 0.312 | 0.688 | 0.101 |   
## | 0.034 | 0.979 | |   
## | 0.032 | 0.070 | |   
## -----------------|-----------|-----------|-----------|  
## Column Total | 1858 | 142 | 2000 |   
## | 0.929 | 0.071 | |   
## -----------------|-----------|-----------|-----------|  
##   
##

#Question 4:Classifying the given customer with best k  
set.seed(123)  
best\_k\_grid<-expand.grid(k=c(best\_k))  
best\_k\_model<-train(Personal.Loan~.,data=trainset\_norm,method="knn",tuneGrid=best\_k\_grid)

## Warning in train.default(x, y, weights = w, ...): You are trying to do  
## regression and your outcome only has two possible values Are you trying to do  
## classification? If so, use a 2 level factor as your outcome column.

best\_k\_model

## k-Nearest Neighbors   
##   
## 3000 samples  
## 13 predictor  
##   
## No pre-processing  
## Resampling: Bootstrapped (25 reps)   
## Summary of sample sizes: 3000, 3000, 3000, 3000, 3000, 3000, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.1954729 0.563781 0.05806894  
##   
## Tuning parameter 'k' was held constant at a value of 6

best\_k\_test<-predict(best\_k\_model,testset\_norm)  
best\_k\_test

## [1] 0

#Question5:confusion matrix for validation and training sets #dividing dataset into traning, validation and testing set

#Question5:confusion matrix for validation and training sets  
#dividing dataset into traning, validation and testing set  
set.seed(123)  
training\_1<-createDataPartition(dataset\_universal\_dummy$CD.Account ,p=0.50,list = FALSE)  
train\_set\_2<-dataset\_universal\_dummy[training\_1,]  
middleset<-dataset\_universal\_dummy[-training\_1,]  
nrow(middleset)

## [1] 2500

training\_2<-createDataPartition(middleset$CD.Account,p=0.6,list = FALSE)  
validation\_set\_2<-middleset[training\_2,]  
test\_set\_2<-middleset[-training\_2,]  
  
nrow(train\_set\_2)

## [1] 2500

nrow(validation\_set\_2)

## [1] 1500

nrow(test\_set\_2)

## [1] 1000

head(train\_set\_2)

## Age Experience Income Family CCAvg Mortgage Personal.Loan Securities.Account  
## 1 25 1 49 4 1.6 0 0 1  
## 2 45 19 34 3 1.5 0 0 1  
## 5 35 8 45 4 1.0 0 0 0  
## 7 53 27 72 2 1.5 0 0 0  
## 11 65 39 105 4 2.4 0 0 0  
## 13 48 23 114 2 3.8 0 0 1  
## CD.Account Online CreditCard education1 education2 education3  
## 1 0 0 0 1 0 0  
## 2 0 0 0 1 0 0  
## 5 0 0 1 0 1 0  
## 7 0 1 0 0 1 0  
## 11 0 0 0 0 0 1  
## 13 0 0 0 0 0 1

#normalizing trainset\_2,validationset\_2,testset\_2  
  
normvar<-c('Age',"Experience","Income","Family","CCAvg","Mortgage","Securities.Account","CD.Account","Online","CreditCard","education1","education2","education3")  
normalization\_values\_2 <- preProcess(train\_set\_2[,normvar],method = c('center','scale'))  
  
trainset\_norm\_2<-predict(normalization\_values\_2,train\_set\_2)  
summary(trainset\_norm\_2)

## Age Experience Income Family   
## Min. :-1.96620 Min. :-2.028011 Min. :-1.4337 Min. :-1.2072   
## 1st Qu.:-0.91026 1st Qu.:-0.887141 1st Qu.:-0.7553 1st Qu.:-1.2072   
## Median : 0.05769 Median :-0.009548 Median :-0.2081 Median :-0.3442   
## Mean : 0.00000 Mean : 0.000000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.84965 3rd Qu.: 0.868044 3rd Qu.: 0.5360 3rd Qu.: 1.3818   
## Max. : 1.90559 Max. : 2.008915 Max. : 3.2937 Max. : 1.3818   
## CCAvg Mortgage Personal.Loan Securities.Account  
## Min. :-1.1117 Min. :-0.5478 Min. :0.0000 Min. :-0.3443   
## 1st Qu.:-0.7101 1st Qu.:-0.5478 1st Qu.:0.0000 1st Qu.:-0.3443   
## Median :-0.2510 Median :-0.5478 Median :0.0000 Median :-0.3443   
## Mean : 0.0000 Mean : 0.0000 Mean :0.0916 Mean : 0.0000   
## 3rd Qu.: 0.3801 3rd Qu.: 0.4302 3rd Qu.:0.0000 3rd Qu.:-0.3443   
## Max. : 4.6261 Max. : 5.7888 Max. :1.0000 Max. : 2.9035   
## CD.Account Online CreditCard education1   
## Min. :-0.2472 Min. :-1.2255 Min. :-0.6508 Min. :-0.8627   
## 1st Qu.:-0.2472 1st Qu.:-1.2255 1st Qu.:-0.6508 1st Qu.:-0.8627   
## Median :-0.2472 Median : 0.8157 Median :-0.6508 Median :-0.8627   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.2472 3rd Qu.: 0.8157 3rd Qu.: 1.5360 3rd Qu.: 1.1587   
## Max. : 4.0441 Max. : 0.8157 Max. : 1.5360 Max. : 1.1587   
## education2 education3   
## Min. :-0.6235 Min. :-0.6439   
## 1st Qu.:-0.6235 1st Qu.:-0.6439   
## Median :-0.6235 Median :-0.6439   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 1.6032 3rd Qu.: 1.5523   
## Max. : 1.6032 Max. : 1.5523

validationset\_norm\_2<-predict(normalization\_values\_2,validation\_set\_2)  
summary(validationset\_norm\_2)

## Age Experience Income Family   
## Min. :-1.966201 Min. :-1.940252 Min. :-1.43373 Min. :-1.20717   
## 1st Qu.:-0.910258 1st Qu.:-0.887141 1st Qu.:-0.75526 1st Qu.:-1.20717   
## Median :-0.030306 Median :-0.009548 Median :-0.22999 Median :-0.34416   
## Mean : 0.005479 Mean : 0.008589 Mean : 0.00752 Mean :-0.02946   
## 3rd Qu.: 0.849647 3rd Qu.: 0.868044 3rd Qu.: 0.54150 3rd Qu.: 0.51884   
## Max. : 1.905589 Max. : 1.833396 Max. : 2.87786 Max. : 1.38184   
## CCAvg Mortgage Personal.Loan Securities.Account  
## Min. :-1.1117 Min. :-0.54777 Min. :0.00000 Min. :-0.34427   
## 1st Qu.:-0.7101 1st Qu.:-0.54777 1st Qu.:0.00000 1st Qu.:-0.34427   
## Median :-0.2510 Median :-0.54777 Median :0.00000 Median :-0.34427   
## Mean :-0.0057 Mean : 0.04763 Mean :0.09533 Mean :-0.00433   
## 3rd Qu.: 0.3227 3rd Qu.: 0.48006 3rd Qu.:0.00000 3rd Qu.:-0.34427   
## Max. : 4.0524 Max. : 5.44953 Max. :1.00000 Max. : 2.90355   
## CD.Account Online CreditCard education1   
## Min. :-0.2472 Min. :-1.22552 Min. :-0.650785 Min. :-0.862725   
## 1st Qu.:-0.2472 1st Qu.:-1.22552 1st Qu.:-0.650785 1st Qu.:-0.862725   
## Median :-0.2472 Median : 0.81565 Median :-0.650785 Median :-0.862725   
## Mean : 0.0246 Mean :-0.02803 Mean : 0.009622 Mean :-0.009703   
## 3rd Qu.:-0.2472 3rd Qu.: 0.81565 3rd Qu.: 1.535992 3rd Qu.: 1.158655   
## Max. : 4.0441 Max. : 0.81565 Max. : 1.535992 Max. : 1.158655   
## education2 education3   
## Min. :-0.62348 Min. :-0.64394   
## 1st Qu.:-0.62348 1st Qu.:-0.64394   
## Median :-0.62348 Median :-0.64394   
## Mean :-0.02078 Mean : 0.03104   
## 3rd Qu.: 1.60325 3rd Qu.: 1.55231   
## Max. : 1.60325 Max. : 1.55231

testset\_norm\_2<-predict(normalization\_values\_2,test\_set\_2)  
summary(testset\_norm\_2)

## Age Experience Income Family   
## Min. :-1.96620 Min. :-2.028011 Min. :-1.43373 Min. :-1.20717   
## 1st Qu.:-0.91026 1st Qu.:-0.887141 1st Qu.:-0.77715 1st Qu.:-1.20717   
## Median :-0.03031 Median :-0.009548 Median :-0.22999 Median :-0.34416   
## Mean :-0.01086 Mean :-0.014726 Mean : 0.01781 Mean : 0.03383   
## 3rd Qu.: 0.84965 3rd Qu.: 0.780285 3rd Qu.: 0.60169 3rd Qu.: 0.51884   
## Max. : 1.90559 Max. : 2.008915 Max. : 2.83409 Max. : 1.38184   
## CCAvg Mortgage Personal.Loan Securities.Account  
## Min. :-1.111724 Min. :-0.547768 Min. :0.000 Min. :-0.34427   
## 1st Qu.:-0.727287 1st Qu.:-0.547768 1st Qu.:0.000 1st Qu.:-0.34427   
## Median :-0.193665 Median :-0.547768 Median :0.000 Median :-0.34427   
## Mean : 0.009749 Mean : 0.008684 Mean :0.108 Mean :-0.01949   
## 3rd Qu.: 0.322744 3rd Qu.: 0.460098 3rd Qu.:0.000 3rd Qu.:-0.34427   
## Max. : 4.626149 Max. : 5.329786 Max. :1.000 Max. : 2.90355   
## CD.Account Online CreditCard education1   
## Min. :-0.24718 Min. :-1.225521 Min. :-0.65078 Min. :-0.86272   
## 1st Qu.:-0.24718 1st Qu.:-1.225521 1st Qu.:-0.65078 1st Qu.:-0.86272   
## Median :-0.24718 Median : 0.815653 Median :-0.65078 Median :-0.86272   
## Mean : 0.02317 Mean : 0.005307 Mean :-0.05379 Mean :-0.06226   
## 3rd Qu.:-0.24718 3rd Qu.: 0.815653 3rd Qu.: 1.53599 3rd Qu.: 1.15865   
## Max. : 4.04408 Max. : 0.815653 Max. : 1.53599 Max. : 1.15865   
## education2 education3   
## Min. :-0.62348 Min. :-0.64394   
## 1st Qu.:-0.62348 1st Qu.:-0.64394   
## Median :-0.62348 Median :-0.64394   
## Mean : 0.03785 Mean : 0.03031   
## 3rd Qu.: 1.60325 3rd Qu.: 1.55231   
## Max. : 1.60325 Max. : 1.55231

#confusion matrix  
library(gmodels)  
  
train\_label\_2<-trainset\_norm\_2[,7]  
validation\_label\_2<-validationset\_norm\_2[,7]  
testing\_label\_2<-testset\_norm\_2[,7]  
  
predicted\_validation\_label\_2<-knn(trainset\_norm\_2,validationset\_norm\_2,cl=train\_label\_2,k=best\_k)  
  
predicted\_test\_label\_2<-knn(trainset\_norm\_2,testset\_norm\_2,cl=train\_label\_2,k=best\_k)  
  
confusionmatrix\_1<-CrossTable(x=validation\_label\_2,y=predicted\_validation\_label\_2,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1500   
##   
##   
## | predicted\_validation\_label\_2   
## validation\_label\_2 | 0 | 1 | Row Total |   
## -------------------|-----------|-----------|-----------|  
## 0 | 1353 | 4 | 1357 |   
## | 0.997 | 0.003 | 0.905 |   
## | 0.964 | 0.041 | |   
## | 0.902 | 0.003 | |   
## -------------------|-----------|-----------|-----------|  
## 1 | 50 | 93 | 143 |   
## | 0.350 | 0.650 | 0.095 |   
## | 0.036 | 0.959 | |   
## | 0.033 | 0.062 | |   
## -------------------|-----------|-----------|-----------|  
## Column Total | 1403 | 97 | 1500 |   
## | 0.935 | 0.065 | |   
## -------------------|-----------|-----------|-----------|  
##   
##

confusionmatrix\_2<-CrossTable(x=testing\_label\_2,y=predicted\_test\_label\_2,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1000   
##   
##   
## | predicted\_test\_label\_2   
## testing\_label\_2 | 0 | 1 | Row Total |   
## ----------------|-----------|-----------|-----------|  
## 0 | 892 | 0 | 892 |   
## | 1.000 | 0.000 | 0.892 |   
## | 0.962 | 0.000 | |   
## | 0.892 | 0.000 | |   
## ----------------|-----------|-----------|-----------|  
## 1 | 35 | 73 | 108 |   
## | 0.324 | 0.676 | 0.108 |   
## | 0.038 | 1.000 | |   
## | 0.035 | 0.073 | |   
## ----------------|-----------|-----------|-----------|  
## Column Total | 927 | 73 | 1000 |   
## | 0.927 | 0.073 | |   
## ----------------|-----------|-----------|-----------|  
##   
##

validation\_table<-table(validation\_label\_2,predicted\_validation\_label\_2)  
confusionMatrix(validation\_table)

## Confusion Matrix and Statistics  
##   
## predicted\_validation\_label\_2  
## validation\_label\_2 0 1  
## 0 1353 4  
## 1 50 93  
##   
## Accuracy : 0.964   
## 95% CI : (0.9533, 0.9728)  
## No Information Rate : 0.9353   
## P-Value [Acc > NIR] : 7.032e-07   
##   
## Kappa : 0.7562   
##   
## Mcnemar's Test P-Value : 9.141e-10   
##   
## Sensitivity : 0.9644   
## Specificity : 0.9588   
## Pos Pred Value : 0.9971   
## Neg Pred Value : 0.6503   
## Prevalence : 0.9353   
## Detection Rate : 0.9020   
## Detection Prevalence : 0.9047   
## Balanced Accuracy : 0.9616   
##   
## 'Positive' Class : 0   
##

test\_table<-table(testing\_label\_2,predicted\_test\_label\_2)  
confusionMatrix(test\_table)

## Confusion Matrix and Statistics  
##   
## predicted\_test\_label\_2  
## testing\_label\_2 0 1  
## 0 892 0  
## 1 35 73  
##   
## Accuracy : 0.965   
## 95% CI : (0.9517, 0.9755)  
## No Information Rate : 0.927   
## P-Value [Acc > NIR] : 2.703e-07   
##   
## Kappa : 0.7882   
##   
## Mcnemar's Test P-Value : 9.081e-09   
##   
## Sensitivity : 0.9622   
## Specificity : 1.0000   
## Pos Pred Value : 1.0000   
## Neg Pred Value : 0.6759   
## Prevalence : 0.9270   
## Detection Rate : 0.8920   
## Detection Prevalence : 0.8920   
## Balanced Accuracy : 0.9811   
##   
## 'Positive' Class : 0   
##

# on comparing the confusion matrix of validation set and testing set it can be seen that accuracy and sensitivity of validation is slightly greater than test set