Assignment 2 FUNDAMENTALS OF MACHINE LEARNING

Brian Yim

2025-09-27

#### 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?

library(readr)  
library(class)  
library(caret)

## Warning: package 'caret' was built under R version 4.4.2

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.4.3

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 4.4.2

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.4.3

##   
## 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("gmodels")

## Warning: package 'gmodels' was built under R version 4.4.3

#reading the data and creating the working dataset

UniversalBank <- read.csv("C:/Users/hyim/OneDrive - Kent State University/Desktop/64060 Assignment2/UniversalBank.csv")  
mydata=data.frame(UniversalBank, header=T, stringsAsFactors=TRUE) # Create the working data set, converting data to factor  
View(mydata)

#check the characteristics and quality of the data

table(is.na(mydata)) #Find if any missing data

##   
## FALSE   
## 75000

data.frame(colnames(mydata)) #Returns column index in the original data

## colnames.mydata.  
## 1 ID  
## 2 Age  
## 3 Experience  
## 4 Income  
## 5 ZIP.Code  
## 6 Family  
## 7 CCAvg  
## 8 Education  
## 9 Mortgage  
## 10 Personal.Loan  
## 11 Securities.Account  
## 12 CD.Account  
## 13 Online  
## 14 CreditCard  
## 15 header

summary(mydata)

## 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 header   
## Min. :0.000 Mode:logical   
## 1st Qu.:0.000 TRUE:5000   
## Median :0.000   
## Mean :0.294   
## 3rd Qu.:1.000   
## Max. :1.000

#manipulate the data

mydata<-mydata[c(-1,-5)] #Remove ID and Zip code  
data.frame(colnames(mydata)) #Returns column index after removing two columns

## colnames.mydata.  
## 1 Age  
## 2 Experience  
## 3 Income  
## 4 Family  
## 5 CCAvg  
## 6 Education  
## 7 Mortgage  
## 8 Personal.Loan  
## 9 Securities.Account  
## 10 CD.Account  
## 11 Online  
## 12 CreditCard  
## 13 header

#dummy Variables (Education\_1, Education\_2, Education\_3) for categorial variable (Education)

mydata$Education\_1<-mydata$Education  
mydata$Education\_1[mydata$Education==1]<-1  
mydata$Education\_1[mydata$Education==2|mydata$Education==3]<-0  
  
mydata$Education\_2<-mydata$Education  
mydata$Education\_2[mydata$Education==2]<-1  
mydata$Education\_2[mydata$Education==1|mydata$Education==3]<-0  
  
mydata$Education\_3<-mydata$Education  
mydata$Education\_3[mydata$Education==3]<-1  
mydata$Education\_3[mydata$Education==1|mydata$Education==2]<-0  
  
mydata<-mydata[,-6] #Remove Education column to count double effect  
  
#Convert the "Personal Loan" Success class as Yes and Fail Class for readability   
mydata$Personal.Loan [mydata$Personal.Loan==1]<-"Success"  
mydata$Personal.Loan [mydata$Personal.Loan==0]<-"Fail"  
  
#Move the Label column (the goal of the research), Personal Loan, to a first position.   
mydata<-mydata%>%relocate(c(Personal.Loan))  
data.frame(colnames(mydata)) #Returns column index to check the result

## colnames.mydata.  
## 1 Personal.Loan  
## 2 Age  
## 3 Experience  
## 4 Income  
## 5 Family  
## 6 CCAvg  
## 7 Mortgage  
## 8 Securities.Account  
## 9 CD.Account  
## 10 Online  
## 11 CreditCard  
## 12 header  
## 13 Education\_1  
## 14 Education\_2  
## 15 Education\_3

str(mydata)

## 'data.frame': 5000 obs. of 15 variables:  
## $ Personal.Loan : chr "Fail" "Fail" "Fail" "Fail" ...  
## $ Age : int 25 45 39 35 35 37 53 50 35 34 ...  
## $ Experience : int 1 19 15 9 8 13 27 24 10 9 ...  
## $ Income : int 49 34 11 100 45 29 72 22 81 180 ...  
## $ Family : int 4 3 1 1 4 4 2 1 3 1 ...  
## $ CCAvg : num 1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...  
## $ Mortgage : int 0 0 0 0 0 155 0 0 104 0 ...  
## $ Securities.Account: int 1 1 0 0 0 0 0 0 0 0 ...  
## $ CD.Account : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Online : int 0 0 0 0 0 1 1 0 1 0 ...  
## $ CreditCard : int 0 0 0 0 1 0 0 1 0 0 ...  
## $ header : logi TRUE TRUE TRUE TRUE TRUE TRUE ...  
## $ Education\_1 : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ Education\_2 : num 0 0 0 1 1 1 1 0 1 0 ...  
## $ Education\_3 : num 0 0 0 0 0 0 0 1 0 1 ...

#dividing the data into training (60%) and validation (40%) data sets

set.seed(200)  
Train\_Index\_q1=createDataPartition(mydata$Personal.Loan, p=0.6, list=FALSE) #Create an index for the training sample (60% of the dataset)  
Train\_Data\_q1=mydata[Train\_Index\_q1,] #Create the training dataset using the index of the training sample  
Validation\_Data\_q1=mydata[-Train\_Index\_q1,] #Create the validation dataset using the reverse index of the training sample (40% of the dataset)   
#Summarize both datasets to ensure the summary statistics are similar.  
summary(Train\_Data\_q1)

## Personal.Loan Age Experience Income   
## Length:3000 Min. :23.00 Min. :-3.00 Min. : 8.00   
## Class :character 1st Qu.:35.00 1st Qu.:10.00 1st Qu.: 39.00   
## Mode :character Median :45.00 Median :20.00 Median : 63.00   
## Mean :45.24 Mean :19.99 Mean : 73.31   
## 3rd Qu.:55.00 3rd Qu.:30.00 3rd Qu.: 98.00   
## Max. :67.00 Max. :43.00 Max. :224.00   
## Family CCAvg Mortgage Securities.Account  
## Min. :1.000 Min. : 0.000 Min. : 0.00 Min. :0.0000   
## 1st Qu.:1.000 1st Qu.: 0.700 1st Qu.: 0.00 1st Qu.:0.0000   
## Median :2.000 Median : 1.500 Median : 0.00 Median :0.0000   
## Mean :2.412 Mean : 1.934 Mean : 55.34 Mean :0.1007   
## 3rd Qu.:3.000 3rd Qu.: 2.500 3rd Qu.: 98.25 3rd Qu.:0.0000   
## Max. :4.000 Max. :10.000 Max. :617.00 Max. :1.0000   
## CD.Account Online CreditCard header   
## Min. :0.00000 Min. :0.000 Min. :0.0000 Mode:logical   
## 1st Qu.:0.00000 1st Qu.:0.000 1st Qu.:0.0000 TRUE:3000   
## Median :0.00000 Median :1.000 Median :0.0000   
## Mean :0.05667 Mean :0.594 Mean :0.2927   
## 3rd Qu.:0.00000 3rd Qu.:1.000 3rd Qu.:1.0000   
## Max. :1.00000 Max. :1.000 Max. :1.0000   
## Education\_1 Education\_2 Education\_3   
## Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.000 Median :0.0000 Median :0.0000   
## Mean :0.411 Mean :0.2827 Mean :0.3063   
## 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.000 Max. :1.0000 Max. :1.0000

summary(Validation\_Data\_q1)

## Personal.Loan Age Experience Income   
## Length:2000 Min. :23.00 Min. :-3.00 Min. : 8.00   
## Class :character 1st Qu.:36.00 1st Qu.:10.00 1st Qu.: 39.00   
## Mode :character Median :46.00 Median :20.00 Median : 64.00   
## Mean :45.49 Mean :20.28 Mean : 74.47   
## 3rd Qu.:55.00 3rd Qu.:30.00 3rd Qu.:101.00   
## Max. :67.00 Max. :43.00 Max. :205.00   
## Family CCAvg Mortgage Securities.Account  
## Min. :1.000 Min. : 0.000 Min. : 0.00 Min. :0.00   
## 1st Qu.:1.000 1st Qu.: 0.700 1st Qu.: 0.00 1st Qu.:0.00   
## Median :2.000 Median : 1.600 Median : 0.00 Median :0.00   
## Mean :2.373 Mean : 1.943 Mean : 58.23 Mean :0.11   
## 3rd Qu.:3.000 3rd Qu.: 2.600 3rd Qu.:103.00 3rd Qu.:0.00   
## Max. :4.000 Max. :10.000 Max. :635.00 Max. :1.00   
## CD.Account Online CreditCard header   
## Min. :0.000 Min. :0.000 Min. :0.000 Mode:logical   
## 1st Qu.:0.000 1st Qu.:0.000 1st Qu.:0.000 TRUE:2000   
## Median :0.000 Median :1.000 Median :0.000   
## Mean :0.066 Mean :0.601 Mean :0.296   
## 3rd Qu.:0.000 3rd Qu.:1.000 3rd Qu.:1.000   
## Max. :1.000 Max. :1.000 Max. :1.000   
## Education\_1 Education\_2 Education\_3   
## Min. :0.0000 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000   
## Median :0.0000 Median :0.0000 Median :0.000   
## Mean :0.4315 Mean :0.2775 Mean :0.291   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :1.0000 Max. :1.0000 Max. :1.000

#data Normalization on the Train set and apply it to the Validation set.(z-score scaling using “center” and “scale” as input method parameters.)

#Copy the original data  
Train\_Norm\_q1<-Train\_Data\_q1  
Validation\_Norm\_q1<-Validation\_Data\_q1  
  
# Normalizing the data sets  
normq1<-preProcess(Train\_Data\_q1,method=c("center", "scale"))  
Train\_Norm\_q1<-predict(normq1, Train\_Data\_q1)  
Validation\_Norm\_q1<-predict(normq1, Validation\_Data\_q1)  
  
# Check the mean and variance of variables in the Training set (expected 0 and 1, respectively)  
summary(Train\_Norm\_q1)

## Personal.Loan Age Experience Income   
## Length:3000 Min. :-1.9326 Min. :-2.002298 Min. :-1.4231   
## Class :character 1st Qu.:-0.8898 1st Qu.:-0.869990 1st Qu.:-0.7477   
## Mode :character Median :-0.0208 Median : 0.001016 Median :-0.2247   
## Mean : 0.0000 Mean : 0.000000 Mean : 0.0000   
## 3rd Qu.: 0.8482 3rd Qu.: 0.872022 3rd Qu.: 0.5379   
## Max. : 1.8910 Max. : 2.004330 Max. : 3.2833   
## Family CCAvg Mortgage Securities.Account  
## Min. :-1.2343 Min. :-1.0966 Min. :-0.5496 Min. :-0.3345   
## 1st Qu.:-1.2343 1st Qu.:-0.6998 1st Qu.:-0.5496 1st Qu.:-0.3345   
## Median :-0.3600 Median :-0.2462 Median :-0.5496 Median :-0.3345   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.5144 3rd Qu.: 0.3207 3rd Qu.: 0.4260 3rd Qu.:-0.3345   
## Max. : 1.3888 Max. : 4.5728 Max. : 5.5773 Max. : 2.9884   
## CD.Account Online CreditCard header   
## Min. :-0.2451 Min. :-1.2094 Min. :-0.6431 Mode:logical   
## 1st Qu.:-0.2451 1st Qu.:-1.2094 1st Qu.:-0.6431 TRUE:3000   
## Median :-0.2451 Median : 0.8266 Median :-0.6431   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.2451 3rd Qu.: 0.8266 3rd Qu.: 1.5544   
## Max. : 4.0794 Max. : 0.8266 Max. : 1.5544   
## Education\_1 Education\_2 Education\_3   
## Min. :-0.8352 Min. :-0.6276 Min. :-0.6644   
## 1st Qu.:-0.8352 1st Qu.:-0.6276 1st Qu.:-0.6644   
## Median :-0.8352 Median :-0.6276 Median :-0.6644   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 1.1969 3rd Qu.: 1.5928 3rd Qu.: 1.5045   
## Max. : 1.1969 Max. : 1.5928 Max. : 1.5045

var(Train\_Norm\_q1[,2:14])

## Age Experience Income Family  
## Age 1.000000000 0.994051331 -0.0541805855 -0.0459752949  
## Experience 0.994051331 1.000000000 -0.0469553336 -0.0518714009  
## Income -0.054180585 -0.046955334 1.0000000000 -0.1386275054  
## Family -0.045975295 -0.051871401 -0.1386275054 1.0000000000  
## CCAvg -0.053978539 -0.051819986 0.6520009769 -0.0830325158  
## Mortgage 0.004347544 0.007189289 0.1934238425 -0.0287929211  
## Securities.Account -0.003974638 -0.005740051 0.0009240143 0.0345647885  
## CD.Account 0.011692905 0.013562278 0.1505197849 0.0063251619  
## Online 0.023451183 0.024172211 0.0144823824 0.0079602253  
## CreditCard 0.016101477 0.017311421 -0.0174571468 0.0003566558  
## header 0.000000000 0.000000000 0.0000000000 0.0000000000  
## Education\_1 -0.022970492 0.001557223 0.2135701280 -0.1135104295  
## Education\_2 -0.036863586 -0.037216080 -0.1277834990 0.1358880445  
## CCAvg Mortgage Securities.Account CD.Account  
## Age -5.397854e-02 0.004347544 -0.0039746381 0.011692905  
## Experience -5.181999e-02 0.007189289 -0.0057400513 0.013562278  
## Income 6.520010e-01 0.193423842 0.0009240143 0.150519785  
## Family -8.303252e-02 -0.028792921 0.0345647885 0.006325162  
## CCAvg 1.000000e+00 0.109962552 0.0145560086 0.128725410  
## Mortgage 1.099626e-01 1.000000000 0.0048281073 0.082123432  
## Securities.Account 1.455601e-02 0.004828107 1.0000000000 0.306118348  
## CD.Account 1.287254e-01 0.082123432 0.3061183482 1.000000000  
## Online -9.722329e-05 -0.018287887 0.0058924036 0.173270763  
## CreditCard -2.136349e-02 -0.035354423 -0.0131126193 0.260616516  
## header 0.000000e+00 0.000000000 0.0000000000 0.000000000  
## Education\_1 1.546284e-01 0.037152766 0.0019769310 -0.020130775  
## Education\_2 -9.500154e-02 -0.018332665 0.0114023996 0.003030970  
## Online CreditCard header Education\_1 Education\_2  
## Age 2.345118e-02 0.0161014774 0 -0.022970492 -0.03686359  
## Experience 2.417221e-02 0.0173114212 0 0.001557223 -0.03721608  
## Income 1.448238e-02 -0.0174571468 0 0.213570128 -0.12778350  
## Family 7.960225e-03 0.0003566558 0 -0.113510429 0.13588804  
## CCAvg -9.722329e-05 -0.0213634901 0 0.154628408 -0.09500154  
## Mortgage -1.828789e-02 -0.0353544227 0 0.037152766 -0.01833267  
## Securities.Account 5.892404e-03 -0.0131126193 0 0.001976931 0.01140240  
## CD.Account 1.732708e-01 0.2606165157 0 -0.020130775 0.00303097  
## Online 1.000000e+00 -0.0067610433 0 0.002204554 0.03510402  
## CreditCard -6.761043e-03 1.0000000000 0 0.018079721 -0.01005691  
## header 0.000000e+00 0.0000000000 0 0.000000000 0.00000000  
## Education\_1 2.204554e-03 0.0180797209 0 1.000000000 -0.52437268  
## Education\_2 3.510402e-02 -0.0100569065 0 -0.524372679 1.00000000

# Check the mean and variance of variables in the other sets (expected none-0 and none-1, respectively)  
summary(Validation\_Norm\_q1)

## Personal.Loan Age Experience Income   
## Length:2000 Min. :-1.93262 Min. :-2.002298 Min. :-1.42313   
## Class :character 1st Qu.:-0.80291 1st Qu.:-0.869990 1st Qu.:-0.74767   
## Mode :character Median : 0.06610 Median : 0.001016 Median :-0.20294   
## Mean : 0.02152 Mean : 0.025317 Mean : 0.02509   
## 3rd Qu.: 0.84821 3rd Qu.: 0.872022 3rd Qu.: 0.60326   
## Max. : 1.89102 Max. : 2.004330 Max. : 2.86934   
## Family CCAvg Mortgage Securities.Account  
## Min. :-1.23434 Min. :-1.096645 Min. :-0.54958 Min. :-0.33451   
## 1st Qu.:-1.23434 1st Qu.:-0.699782 1st Qu.:-0.54958 1st Qu.:-0.33451   
## Median :-0.35995 Median :-0.189530 Median :-0.54958 Median :-0.33451   
## Mean :-0.03337 Mean : 0.005156 Mean : 0.02864 Mean : 0.03101   
## 3rd Qu.: 0.51443 3rd Qu.: 0.377416 3rd Qu.: 0.47322 3rd Qu.:-0.33451   
## Max. : 1.38881 Max. : 4.572820 Max. : 5.75602 Max. : 2.98844   
## CD.Account Online CreditCard header   
## Min. :-0.24505 Min. :-1.20937 Min. :-0.643135 Mode:logical   
## 1st Qu.:-0.24505 1st Qu.:-1.20937 1st Qu.:-0.643135 TRUE:2000   
## Median :-0.24505 Median : 0.82660 Median :-0.643135   
## Mean : 0.04036 Mean : 0.01425 Mean : 0.007325   
## 3rd Qu.:-0.24505 3rd Qu.: 0.82660 3rd Qu.: 1.554365   
## Max. : 4.07940 Max. : 0.82660 Max. : 1.554365   
## Education\_1 Education\_2 Education\_3   
## Min. :-0.83520 Min. :-0.62763 Min. :-0.66443   
## 1st Qu.:-0.83520 1st Qu.:-0.62763 1st Qu.:-0.66443   
## Median :-0.83520 Median :-0.62763 Median :-0.66443   
## Mean : 0.04166 Mean :-0.01147 Mean :-0.03326   
## 3rd Qu.: 1.19692 3rd Qu.: 1.59276 3rd Qu.: 1.50455   
## Max. : 1.19692 Max. : 1.59276 Max. : 1.50455

var(Validation\_Norm\_q1[,2:14])

## Age Experience Income Family  
## Age 0.981047054 0.982270845 -0.057140323 -0.046631861  
## Experience 0.982270845 0.994445895 -0.046628457 -0.053430104  
## Income -0.057140323 -0.046628457 1.015349202 -0.187970490  
## Family -0.046631861 -0.053430104 -0.187970490 1.017352536  
## CCAvg -0.047466890 -0.046275088 0.627252312 -0.147036171  
## Mortgage -0.038441054 -0.037917850 0.233328011 -0.008050429  
## Securities.Account 0.004457912 0.005012564 -0.008522668 -0.000247094  
## CD.Account 0.002578516 0.005685173 0.212239247 0.027802146  
## Online -0.001272214 -0.001805988 0.013649764 0.014296063  
## CreditCard -0.005092135 -0.003654574 0.020089778 0.028728613  
## header 0.000000000 0.000000000 0.000000000 0.000000000  
## Education\_1 -0.035449396 -0.010251319 0.227365590 -0.127402090  
## Education\_2 0.015030431 0.010657183 -0.129363817 0.144428055  
## CCAvg Mortgage Securities.Account CD.Account  
## Age -0.047466890 -0.038441054 0.004457912 0.002578516  
## Experience -0.046275088 -0.037917850 0.005012564 0.005685173  
## Income 0.627252312 0.233328011 -0.008522668 0.212239247  
## Family -0.147036171 -0.008050429 -0.000247094 0.027802146  
## CCAvg 0.954826557 0.109993906 0.016052674 0.155311442  
## Mortgage 0.109993906 1.050401616 -0.021665069 0.108519780  
## Securities.Account 0.016052674 -0.021665069 1.081555289 0.370067837  
## CD.Account 0.155311442 0.108519780 0.370067837 1.153374584  
## Online -0.008835162 0.012067097 0.022946301 0.192332668  
## CreditCard 0.015431624 0.034627438 -0.018702966 0.327674751  
## header 0.000000000 0.000000000 0.000000000 0.000000000  
## Education\_1 0.157945887 0.052057212 0.013748492 -0.008607561  
## Education\_2 -0.080835678 -0.052451027 -0.003875527 0.011384043  
## Online CreditCard header Education\_1 Education\_2  
## Age -0.001272214 -0.005092135 0 -0.035449396 0.015030431  
## Experience -0.001805988 -0.003654574 0 -0.010251319 0.010657183  
## Income 0.013649764 0.020089778 0 0.227365590 -0.129363817  
## Family 0.014296063 0.028728613 0 -0.127402090 0.144428055  
## CCAvg -0.008835162 0.015431624 0 0.157945887 -0.080835678  
## Mortgage 0.012067097 0.034627438 0 0.052057212 -0.052451027  
## Securities.Account 0.022946301 -0.018702966 0 0.013748492 -0.003875527  
## CD.Account 0.192332668 0.327674751 0 -0.008607561 0.011384043  
## Online 0.994505757 0.020608804 0 0.004836891 -0.001255109  
## CreditCard 0.020608804 1.006791167 0 0.010168746 -0.015328708  
## header 0.000000000 0.000000000 0 0.000000000 0.000000000  
## Education\_1 0.004836891 0.010168746 0 1.013506562 -0.540554935  
## Education\_2 -0.001255109 -0.015328708 0 -0.540554935 0.988957530

#modeling k-NN and define predictors and labels for modeling on Train and Validation sets

Train\_Predictors<- Train\_Norm\_q1[,2:14]  
Train\_Label <- Train\_Norm\_q1[,1]  
Validation\_Predictors<-Validation\_Norm\_q1[,2:14]  
Validation\_Label<- Validation\_Norm\_q1[,1]

#train a knn model

Pred.Model<-knn(Train\_Predictors, Validation\_Predictors, cl=Train\_Label,k=1)

#confusion matrix for the validation set

CrossTable(x=Validation\_Label,y= Pred.Model, prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 2000   
##   
##   
## | Pred.Model   
## Validation\_Label | Fail | Success | Row Total |   
## -----------------|-----------|-----------|-----------|  
## Fail | 1780 | 28 | 1808 |   
## | 0.985 | 0.015 | 0.904 |   
## | 0.965 | 0.179 | |   
## | 0.890 | 0.014 | |   
## -----------------|-----------|-----------|-----------|  
## Success | 64 | 128 | 192 |   
## | 0.333 | 0.667 | 0.096 |   
## | 0.035 | 0.821 | |   
## | 0.032 | 0.064 | |   
## -----------------|-----------|-----------|-----------|  
## Column Total | 1844 | 156 | 2000 |   
## | 0.922 | 0.078 | |   
## -----------------|-----------|-----------|-----------|  
##   
##

Validation\_Label<-as.factor(Validation\_Label)  
confusionMatrix(Pred.Model,Validation\_Label)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Fail Success  
## Fail 1780 64  
## Success 28 128  
##   
## Accuracy : 0.954   
## 95% CI : (0.9439, 0.9628)  
## No Information Rate : 0.904   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.7107   
##   
## Mcnemar's Test P-Value : 0.0002633   
##   
## Sensitivity : 0.9845   
## Specificity : 0.6667   
## Pos Pred Value : 0.9653   
## Neg Pred Value : 0.8205   
## Prevalence : 0.9040   
## Detection Rate : 0.8900   
## Detection Prevalence : 0.9220   
## Balanced Accuracy : 0.8256   
##   
## 'Positive' Class : Fail   
##

#evaluation of the New Applicant, whose profile was normalized based on the training set, K=1

Applicant = data.frame(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, CreditCard = 1) #New Applicant Profile  
Applicant\_Norm\_q1<-predict(normq1, Applicant) # Normalizing the profile  
Applicant\_Predictors<- Applicant\_Norm\_q1  
  
Pred.Applicant <- knn(Train\_Predictors, Applicant\_Predictors, cl=Train\_Label, k=1, prob=TRUE)  
attributes(Pred.Applicant)

## $levels  
## [1] "Fail" "Success"  
##   
## $class  
## [1] "factor"  
##   
## $prob  
## [1] 1

Result<-data.frame(Pred.Applicant)  
Result

## Pred.Applicant  
## 1 Fail

head(Pred.Applicant)

## [1] Fail  
## Levels: Fail Success

#Answer for the Q1: With the kNN model set at k=1, the customer ended up being labeled as ‘Fail’ for loan acceptance, with a 100% probability

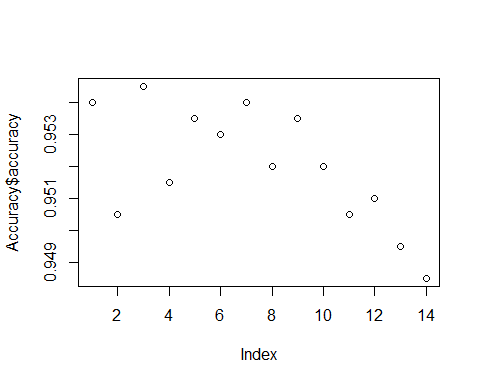
#### 2. What is a choice of k that balances between overfitting and ignoring the predictor information?

#compute knn for different k on Training and Validation

Accuracy<- data.frame(k = seq(1, 14, 1), accuracy = rep(0, 14))#create a table with two columns :k and accuracy  
#To determin k, I am using the performance on the validation set value of k from 1 to 14  
for(i in 1:14) {  
 Pred.Model <- knn(Train\_Predictors, Validation\_Predictors,   
 cl = Train\_Label, k = i)  
 Accuracy[i, 2] <- confusionMatrix(Pred.Model, Validation\_Label)$overall[1]   
}  
Accuracy

## k accuracy  
## 1 1 0.9540  
## 2 2 0.9505  
## 3 3 0.9545  
## 4 4 0.9515  
## 5 5 0.9535  
## 6 6 0.9530  
## 7 7 0.9540  
## 8 8 0.9520  
## 9 9 0.9535  
## 10 10 0.9520  
## 11 11 0.9505  
## 12 12 0.9510  
## 13 13 0.9495  
## 14 14 0.9485

plot(Accuracy$accuracy)



#Answer for Q2: The algorithm worked best with k=3. Using our current data (60% training, 40% validation), k=3 seems to give a good balance between overfitting and underfitting

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

Pred.Model2<-knn(Train\_Predictors, Validation\_Predictors, cl=Train\_Label,k=3)  
CrossTable(x=Validation\_Label,y= Pred.Model2, prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 2000   
##   
##   
## | Pred.Model2   
## Validation\_Label | Fail | Success | Row Total |   
## -----------------|-----------|-----------|-----------|  
## Fail | 1790 | 18 | 1808 |   
## | 0.990 | 0.010 | 0.904 |   
## | 0.961 | 0.131 | |   
## | 0.895 | 0.009 | |   
## -----------------|-----------|-----------|-----------|  
## Success | 73 | 119 | 192 |   
## | 0.380 | 0.620 | 0.096 |   
## | 0.039 | 0.869 | |   
## | 0.036 | 0.059 | |   
## -----------------|-----------|-----------|-----------|  
## Column Total | 1863 | 137 | 2000 |   
## | 0.931 | 0.068 | |   
## -----------------|-----------|-----------|-----------|  
##   
##

confusionMatrix(Pred.Model2,Validation\_Label)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Fail Success  
## Fail 1790 73  
## Success 18 119  
##   
## Accuracy : 0.9545   
## 95% CI : (0.9444, 0.9632)  
## No Information Rate : 0.904   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.6994   
##   
## Mcnemar's Test P-Value : 1.507e-08   
##   
## Sensitivity : 0.9900   
## Specificity : 0.6198   
## Pos Pred Value : 0.9608   
## Neg Pred Value : 0.8686   
## Prevalence : 0.9040   
## Detection Rate : 0.8950   
## Detection Prevalence : 0.9315   
## Balanced Accuracy : 0.8049   
##   
## 'Positive' Class : Fail   
##

#Answer for Q3: The confusion matrix for k=3 showed an accuracy of 0.9545, meaning about 95.45% of the loan acceptance predictions were correct, while roughly 4.55% were misclassified

#### 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.

#best k was 3 with the highest accuracy so

#Evaluation of the New Person using the best k=3  
Pred.Applicant <- knn(Train\_Predictors, Applicant\_Predictors, cl=Train\_Label, k=3, prob=TRUE)  
attributes(Pred.Applicant)

## $levels  
## [1] "Fail" "Success"  
##   
## $class  
## [1] "factor"  
##   
## $prob  
## [1] 1

Result2<-data.frame(Pred.Applicant)  
Result2

## Pred.Applicant  
## 1 Fail

head(Pred.Applicant)

## [1] Fail  
## Levels: Fail Success

#Answer for Q4: With the current kNN model (k=3), the customer was classified as ‘fail’ for loan acceptance, with a probability of about 66.7%

#### 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.

#dividing the data into training (50%), validation (30%), and test (20%) sets

set.seed(1)  
Test\_Index\_q5=createDataPartition(mydata$Personal.Loan, p=0.2, list=FALSE) #Create an index for the testing sample (20% of the dataset)  
Test\_Data\_q5= mydata[Test\_Index\_q5,] #Create the testing dataset using the index of the Test\_Index sample  
TraVal\_Data\_q5= mydata[-Test\_Index\_q5,] #Validation and Training data is rest  
  
Train\_Index\_q5=createDataPartition(TraVal\_Data\_q5$Personal.Loan, p=0.625, list=FALSE) #Create an index for the testing sample (62.5% of TraVal dataset = 50% of the total dataset)  
Train\_Data\_q5=TraVal\_Data\_q5[Train\_Index\_q5,] #Create the training dataset using the index of the Train sample  
Validation\_Data\_q5=TraVal\_Data\_q5 [-Train\_Index\_q5,] #Create the validation dataset using the reverse index of the training sample (37.5% of TraVal dataset = 20% of the total dataset)  
#Summarize both datasets to ensure the summary statistics are similar.  
summary(Train\_Data\_q5)

## Personal.Loan Age Experience Income   
## Length:2500 Min. :23.00 Min. :-3.00 Min. : 8.00   
## Class :character 1st Qu.:35.00 1st Qu.:10.00 1st Qu.: 39.00   
## Mode :character Median :45.00 Median :20.00 Median : 64.00   
## Mean :45.24 Mean :20.01 Mean : 74.66   
## 3rd Qu.:55.00 3rd Qu.:30.00 3rd Qu.: 99.00   
## Max. :67.00 Max. :43.00 Max. :224.00   
## Family CCAvg Mortgage Securities.Account  
## Min. :1.000 Min. :0.000 Min. : 0.00 Min. :0.0000   
## 1st Qu.:1.000 1st Qu.:0.700 1st Qu.: 0.00 1st Qu.:0.0000   
## Median :2.000 Median :1.600 Median : 0.00 Median :0.0000   
## Mean :2.408 Mean :1.954 Mean : 58.23 Mean :0.0996   
## 3rd Qu.:4.000 3rd Qu.:2.600 3rd Qu.:102.00 3rd Qu.:0.0000   
## Max. :4.000 Max. :9.300 Max. :635.00 Max. :1.0000   
## CD.Account Online CreditCard header   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Mode:logical   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 TRUE:2500   
## Median :0.0000 Median :1.0000 Median :0.0000   
## Mean :0.0628 Mean :0.5984 Mean :0.3028   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000   
## Education\_1 Education\_2 Education\_3   
## Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.000 Median :0.0000 Median :0.0000   
## Mean :0.428 Mean :0.2792 Mean :0.2928   
## 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.000 Max. :1.0000 Max. :1.0000

summary(Validation\_Data\_q5)

## Personal.Loan Age Experience Income   
## Length:1500 Min. :23.0 Min. :-3.00 Min. : 8.00   
## Class :character 1st Qu.:35.0 1st Qu.:10.00 1st Qu.: 39.00   
## Mode :character Median :45.0 Median :20.00 Median : 62.00   
## Mean :45.3 Mean :20.08 Mean : 71.93   
## 3rd Qu.:55.0 3rd Qu.:30.00 3rd Qu.: 93.00   
## Max. :67.0 Max. :43.00 Max. :204.00   
## Family CCAvg Mortgage Securities.Account  
## Min. :1.000 Min. : 0.000 Min. : 0.00 Min. :0.000   
## 1st Qu.:1.000 1st Qu.: 0.600 1st Qu.: 0.00 1st Qu.:0.000   
## Median :2.000 Median : 1.500 Median : 0.00 Median :0.000   
## Mean :2.395 Mean : 1.874 Mean : 53.86 Mean :0.098   
## 3rd Qu.:3.000 3rd Qu.: 2.500 3rd Qu.: 95.00 3rd Qu.:0.000   
## Max. :4.000 Max. :10.000 Max. :617.00 Max. :1.000   
## CD.Account Online CreditCard header   
## Min. :0.000 Min. :0.0000 Min. :0.0000 Mode:logical   
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000 TRUE:1500   
## Median :0.000 Median :1.0000 Median :0.0000   
## Mean :0.054 Mean :0.5907 Mean :0.2907   
## 3rd Qu.:0.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.000 Max. :1.0000 Max. :1.0000   
## Education\_1 Education\_2 Education\_3   
## Min. :0.0000 Min. :0.000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000   
## Median :0.0000 Median :0.000 Median :0.0000   
## Mean :0.4133 Mean :0.282 Mean :0.3047   
## 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.000 Max. :1.0000

summary(Test\_Data\_q5)

## Personal.Loan Age Experience Income   
## Length:1000 Min. :23.00 Min. :-3.00 Min. : 8.00   
## Class :character 1st Qu.:35.00 1st Qu.:10.00 1st Qu.: 39.00   
## Mode :character Median :46.00 Median :21.00 Median : 64.00   
## Mean :45.65 Mean :20.38 Mean : 74.33   
## 3rd Qu.:56.00 3rd Qu.:30.00 3rd Qu.:102.00   
## Max. :67.00 Max. :41.00 Max. :205.00   
## Family CCAvg Mortgage Securities.Account  
## Min. :1.000 Min. : 0.000 Min. : 0.00 Min. :0.000   
## 1st Qu.:1.000 1st Qu.: 0.700 1st Qu.: 0.00 1st Qu.:0.000   
## Median :2.000 Median : 1.600 Median : 0.00 Median :0.000   
## Mean :2.371 Mean : 1.993 Mean : 56.12 Mean :0.126   
## 3rd Qu.:3.000 3rd Qu.: 2.600 3rd Qu.:104.00 3rd Qu.:0.000   
## Max. :4.000 Max. :10.000 Max. :582.00 Max. :1.000   
## CD.Account Online CreditCard header Education\_1   
## Min. :0.000 Min. :0.000 Min. :0.000 Mode:logical Min. :0.000   
## 1st Qu.:0.000 1st Qu.:0.000 1st Qu.:0.000 TRUE:1000 1st Qu.:0.000   
## Median :0.000 Median :1.000 Median :0.000 Median :0.000   
## Mean :0.064 Mean :0.602 Mean :0.277 Mean :0.406   
## 3rd Qu.:0.000 3rd Qu.:1.000 3rd Qu.:1.000 3rd Qu.:1.000   
## Max. :1.000 Max. :1.000 Max. :1.000 Max. :1.000   
## Education\_2 Education\_3   
## Min. :0.000 Min. :0.000   
## 1st Qu.:0.000 1st Qu.:0.000   
## Median :0.000 Median :0.000   
## Mean :0.282 Mean :0.312   
## 3rd Qu.:1.000 3rd Qu.:1.000   
## Max. :1.000 Max. :1.000

#data Normalization on the Train set and apply it to Validation set (z-score scaling using “center” and “scale” as input method parameters.)

#Copy the original data  
Train\_Norm\_q5<-Train\_Data\_q5  
Valid\_Norm\_q5<-Validation\_Data\_q5  
TraVal\_Norm\_q5<-TraVal\_Data\_q5  
Test\_Norm\_q5<- Test\_Data\_q5  
  
# use preProcess to normalize all data  
norm.values.q5<-preProcess(Train\_Data\_q5,method=c("center", "scale"))  
Train\_Norm\_q5<-predict(norm.values.q5, Train\_Data\_q5)  
Valid\_Norm\_q5<-predict(norm.values.q5, Validation\_Data\_q5)  
  
# Check the mean and variance of variables in the other sets  
summary(Train\_Norm\_q5)

## Personal.Loan Age Experience Income   
## Length:2500 Min. :-1.95507 Min. :-2.0156196 Min. :-1.4291   
## Class :character 1st Qu.:-0.90021 1st Qu.:-0.8766931 1st Qu.:-0.7645   
## Mode :character Median :-0.02117 Median :-0.0005957 Median :-0.2286   
## Mean : 0.00000 Mean : 0.0000000 Mean : 0.0000   
## 3rd Qu.: 0.85788 3rd Qu.: 0.8755016 3rd Qu.: 0.5218   
## Max. : 1.91273 Max. : 2.0144281 Max. : 3.2016   
## Family CCAvg Mortgage Securities.Account  
## Min. :-1.2176 Min. :-1.1365 Min. :-0.5634 Min. :-0.3325   
## 1st Qu.:-1.2176 1st Qu.:-0.7295 1st Qu.:-0.5634 1st Qu.:-0.3325   
## Median :-0.3526 Median :-0.2061 Median :-0.5634 Median :-0.3325   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 1.3775 3rd Qu.: 0.3754 3rd Qu.: 0.4234 3rd Qu.:-0.3325   
## Max. : 1.3775 Max. : 4.2716 Max. : 5.5801 Max. : 3.0061   
## CD.Account Online CreditCard header   
## Min. :-0.2588 Min. :-1.2204 Min. :-0.6589 Mode:logical   
## 1st Qu.:-0.2588 1st Qu.:-1.2204 1st Qu.:-0.6589 TRUE:2500   
## Median :-0.2588 Median : 0.8191 Median :-0.6589   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.2588 3rd Qu.: 0.8191 3rd Qu.: 1.5171   
## Max. : 3.8623 Max. : 0.8191 Max. : 1.5171   
## Education\_1 Education\_2 Education\_3   
## Min. :-0.8648 Min. :-0.6222 Min. :-0.6433   
## 1st Qu.:-0.8648 1st Qu.:-0.6222 1st Qu.:-0.6433   
## Median :-0.8648 Median :-0.6222 Median :-0.6433   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 1.1558 3rd Qu.: 1.6064 3rd Qu.: 1.5538   
## Max. : 1.1558 Max. : 1.6064 Max. : 1.5538

var(Train\_Norm\_q5[,2:14])

## Age Experience Income Family  
## Age 1.000000000 0.994236439 -0.048046189 -0.060928792  
## Experience 0.994236439 1.000000000 -0.035604313 -0.066381963  
## Income -0.048046189 -0.035604313 1.000000000 -0.177306640  
## Family -0.060928792 -0.066381963 -0.177306640 1.000000000  
## CCAvg -0.037847973 -0.031713568 0.633847548 -0.113907098  
## Mortgage -0.023490974 -0.022277096 0.219004352 -0.042052594  
## Securities.Account 0.007168541 0.006005190 0.007084735 0.002897944  
## CD.Account 0.031485501 0.033075812 0.173915080 0.004289313  
## Online 0.038579481 0.037382190 0.004651307 0.013576094  
## CreditCard 0.007020028 0.006625588 0.000139475 0.016907405  
## header 0.000000000 0.000000000 0.000000000 0.000000000  
## Education\_1 -0.017603009 0.006639420 0.235883965 -0.138584251  
## Education\_2 -0.013647017 -0.016231835 -0.122852072 0.153131335  
## CCAvg Mortgage Securities.Account CD.Account  
## Age -0.037847973 -0.02349097 0.007168541 0.031485501  
## Experience -0.031713568 -0.02227710 0.006005190 0.033075812  
## Income 0.633847548 0.21900435 0.007084735 0.173915080  
## Family -0.113907098 -0.04205259 0.002897944 0.004289313  
## CCAvg 1.000000000 0.11017308 0.020055191 0.139504199  
## Mortgage 0.110173084 1.00000000 -0.020695478 0.056628647  
## Securities.Account 0.020055191 -0.02069548 1.000000000 0.321331082  
## CD.Account 0.139504199 0.05662865 0.321331082 1.000000000  
## Online -0.022757780 -0.01539251 0.024518011 0.178429475  
## CreditCard -0.007074192 -0.03076387 -0.012782982 0.270786188  
## header 0.000000000 0.00000000 0.000000000 0.000000000  
## Education\_1 0.169356644 0.05886753 -0.017741496 -0.017314632  
## Education\_2 -0.086516379 -0.03477219 0.004404275 0.011634694  
## Online CreditCard header Education\_1 Education\_2  
## Age 0.038579481 0.007020028 0 -0.017603009 -0.013647017  
## Experience 0.037382190 0.006625588 0 0.006639420 -0.016231835  
## Income 0.004651307 0.000139475 0 0.235883965 -0.122852072  
## Family 0.013576094 0.016907405 0 -0.138584251 0.153131335  
## CCAvg -0.022757780 -0.007074192 0 0.169356644 -0.086516379  
## Mortgage -0.015392506 -0.030763869 0 0.058867526 -0.034772190  
## Securities.Account 0.024518011 -0.012782982 0 -0.017741496 0.004404275  
## CD.Account 0.178429475 0.270786188 0 -0.017314632 0.011634694  
## Online 1.000000000 -0.001755978 0 0.029208890 0.004213964  
## CreditCard -0.001755978 1.000000000 0 -0.005271392 -0.010390802  
## header 0.000000000 0.000000000 0 0.000000000 0.000000000  
## Education\_1 0.029208890 -0.005271392 0 1.000000000 -0.538361740  
## Education\_2 0.004213964 -0.010390802 0 -0.538361740 1.000000000

summary(Valid\_Norm\_q5)

## Personal.Loan Age Experience Income   
## Length:1500 Min. :-1.955066 Min. :-2.0156196 Min. :-1.42911   
## Class :character 1st Qu.:-0.900212 1st Qu.:-0.8766931 1st Qu.:-0.76452   
## Mode :character Median :-0.021167 Median :-0.0005957 Median :-0.27144   
## Mean : 0.004852 Mean : 0.0067635 Mean :-0.05865   
## 3rd Qu.: 0.857877 3rd Qu.: 0.8755016 3rd Qu.: 0.39315   
## Max. : 1.912732 Max. : 2.0144281 Max. : 2.77282   
## Family CCAvg Mortgage Securities.Account   
## Min. :-1.21762 Min. :-1.1365 Min. :-0.56339 Min. :-0.332526   
## 1st Qu.:-1.21762 1st Qu.:-0.7876 1st Qu.:-0.56339 1st Qu.:-0.332526   
## Median :-0.35259 Median :-0.2642 Median :-0.56339 Median :-0.332526   
## Mean :-0.01119 Mean :-0.0467 Mean :-0.04233 Mean :-0.005342   
## 3rd Qu.: 0.51245 3rd Qu.: 0.3173 3rd Qu.: 0.35571 3rd Qu.:-0.332526   
## Max. : 1.37748 Max. : 4.6787 Max. : 5.40592 Max. : 3.006085   
## CD.Account Online CreditCard header   
## Min. :-0.25881 Min. :-1.22043 Min. :-0.6589 Mode:logical   
## 1st Qu.:-0.25881 1st Qu.:-1.22043 1st Qu.:-0.6589 TRUE:1500   
## Median :-0.25881 Median : 0.81906 Median :-0.6589   
## Mean :-0.03627 Mean :-0.01577 Mean :-0.0264   
## 3rd Qu.:-0.25881 3rd Qu.: 0.81906 3rd Qu.: 1.5171   
## Max. : 3.86233 Max. : 0.81906 Max. : 1.5171   
## Education\_1 Education\_2 Education\_3   
## Min. :-0.86484 Min. :-0.62225 Min. :-0.64332   
## 1st Qu.:-0.86484 1st Qu.:-0.62225 1st Qu.:-0.64332   
## Median :-0.86484 Median :-0.62225 Median :-0.64332   
## Mean :-0.02964 Mean : 0.00624 Mean : 0.02607   
## 3rd Qu.: 1.15582 3rd Qu.: 1.60643 3rd Qu.: 1.55381   
## Max. : 1.15582 Max. : 1.60643 Max. : 1.55381

var(Valid\_Norm\_q5[,2:14])

## Age Experience Income Family  
## Age 1.001796577 0.992745503 -0.051375012 -0.030093057  
## Experience 0.992745503 0.995114121 -0.047924629 -0.037954704  
## Income -0.051375012 -0.047924629 0.926431460 -0.128079895  
## Family -0.030093057 -0.037954704 -0.128079895 0.990564640  
## CCAvg -0.033429611 -0.036432461 0.616094760 -0.092519153  
## Mortgage -0.009787281 -0.007890278 0.181771064 0.011808472  
## Securities.Account -0.002449638 -0.004165562 0.008880013 0.040428222  
## CD.Account -0.016669578 -0.013922774 0.174877002 0.009588890  
## Online 0.006547364 0.006147788 0.037544142 0.003913689  
## CreditCard 0.025258543 0.029679966 0.006521459 -0.003860870  
## header 0.000000000 0.000000000 0.000000000 0.000000000  
## Education\_1 -0.032173958 -0.007922039 0.165860363 -0.071938622  
## Education\_2 -0.010352054 -0.011792363 -0.126744991 0.096530398  
## CCAvg Mortgage Securities.Account CD.Account  
## Age -0.033429611 -0.009787281 -0.002449638 -0.0166695784  
## Experience -0.036432461 -0.007890278 -0.004165562 -0.0139227738  
## Income 0.616094760 0.181771064 0.008880013 0.1748770016  
## Family -0.092519153 0.011808472 0.040428222 0.0095888901  
## CCAvg 1.008403274 0.080155423 0.038686003 0.1095767684  
## Mortgage 0.080155423 0.939721057 0.009133566 0.1023904888  
## Securities.Account 0.038686003 0.009133566 0.985947209 0.2300366138  
## CD.Account 0.109576768 0.102390489 0.230036614 0.8681781395  
## Online 0.019383683 0.026284248 -0.040100232 0.1578726739  
## CreditCard -0.002820973 0.031499716 -0.032606678 0.2659514803  
## header 0.000000000 0.000000000 0.000000000 0.0000000000  
## Education\_1 0.131354062 0.018936936 0.041584288 -0.0304431307  
## Education\_2 -0.102677907 -0.037887053 -0.007217329 0.0009681009  
## Online CreditCard header Education\_1 Education\_2  
## Age 0.006547364 0.025258543 0 -0.032173958 -0.0103520539  
## Experience 0.006147788 0.029679966 0 -0.007922039 -0.0117923634  
## Income 0.037544142 0.006521459 0 0.165860363 -0.1267449909  
## Family 0.003913689 -0.003860870 0 -0.071938622 0.0965303981  
## CCAvg 0.019383683 -0.002820973 0 0.131354062 -0.1026779067  
## Mortgage 0.026284248 0.031499716 0 0.018936936 -0.0378870526  
## Securities.Account -0.040100232 -0.032606678 0 0.041584288 -0.0072173292  
## CD.Account 0.157872674 0.265951480 0 -0.030443131 0.0009681009  
## Online 1.006352569 0.007310634 0 -0.047323521 0.0671585621  
## CreditCard 0.007310634 0.976895757 0 0.046306135 -0.0063151360  
## header 0.000000000 0.000000000 0 0.000000000 0.0000000000  
## Education\_1 -0.047323521 0.046306135 0 0.990758756 -0.5252673635  
## Education\_2 0.067158562 -0.006315136 0 -0.525267363 1.0063735972

#modeling k-NN, Define predictors and labels for modeling on Train and Validation sets

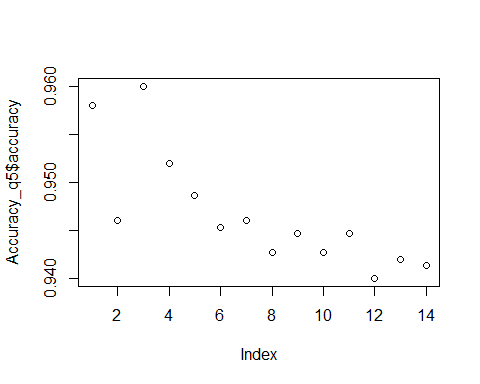
Train\_Predictors\_q5<- Train\_Norm\_q5[,2:14]  
Train\_Label\_q5<- Train\_Norm\_q5[,1]  
Train\_Label\_q5<-as.factor(Train\_Label\_q5)  
Validation\_Predictors\_q5<- Valid\_Norm\_q5[,2:14]  
Validation\_Label\_q5<- Valid\_Norm\_q5[,1]  
Validation\_Label\_q5<-as.factor(Validation\_Label\_q5)

#hypertuning using the validation set. Compute knn for different k on Training and Validation

Accuracy\_q5<- data.frame(k = seq(1, 14, 1), accuracy = rep(0, 14))#create a table with two columns :k and accuracy  
#To determin k, I am using the performance on the validation set value of k from 1 to 14  
for(i in 1:14) {  
 Pred.Model\_k <- knn(Train\_Predictors\_q5, Validation\_Predictors\_q5,   
 cl = Train\_Label\_q5, k = i)  
 Accuracy\_q5[i, 2] <- confusionMatrix(Pred.Model\_k, Validation\_Label\_q5)$overall[1]   
}  
Accuracy\_q5

## k accuracy  
## 1 1 0.9580000  
## 2 2 0.9460000  
## 3 3 0.9600000  
## 4 4 0.9520000  
## 5 5 0.9486667  
## 6 6 0.9453333  
## 7 7 0.9460000  
## 8 8 0.9426667  
## 9 9 0.9446667  
## 10 10 0.9426667  
## 11 11 0.9446667  
## 12 12 0.9400000  
## 13 13 0.9420000  
## 14 14 0.9413333

plot(Accuracy\_q5$accuracy)



#train a knn model for Training set.

Pred.Model.train<-knn(Train\_Predictors\_q5, Train\_Predictors\_q5, cl=Train\_Label\_q5,k=3)

#confusion matrix for the Training set.

CrossTable(x=Train\_Label\_q5,y= Pred.Model.train, prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 2500   
##   
##   
## | Pred.Model.train   
## Train\_Label\_q5 | Fail | Success | Row Total |   
## ---------------|-----------|-----------|-----------|  
## Fail | 2256 | 4 | 2260 |   
## | 0.998 | 0.002 | 0.904 |   
## | 0.973 | 0.022 | |   
## | 0.902 | 0.002 | |   
## ---------------|-----------|-----------|-----------|  
## Success | 63 | 177 | 240 |   
## | 0.263 | 0.738 | 0.096 |   
## | 0.027 | 0.978 | |   
## | 0.025 | 0.071 | |   
## ---------------|-----------|-----------|-----------|  
## Column Total | 2319 | 181 | 2500 |   
## | 0.928 | 0.072 | |   
## ---------------|-----------|-----------|-----------|  
##   
##

confusionMatrix(Pred.Model.train, Train\_Label\_q5)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Fail Success  
## Fail 2256 63  
## Success 4 177  
##   
## Accuracy : 0.9732   
## 95% CI : (0.9661, 0.9792)  
## No Information Rate : 0.904   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.8265   
##   
## Mcnemar's Test P-Value : 1.382e-12   
##   
## Sensitivity : 0.9982   
## Specificity : 0.7375   
## Pos Pred Value : 0.9728   
## Neg Pred Value : 0.9779   
## Prevalence : 0.9040   
## Detection Rate : 0.9024   
## Detection Prevalence : 0.9276   
## Balanced Accuracy : 0.8679   
##   
## 'Positive' Class : Fail   
##

#train a knn model for the Validation set

Pred.Model.validation<-knn(Train\_Predictors\_q5, Validation\_Predictors\_q5, cl=Train\_Label\_q5,k=3)

#confusion matrix for the validation set

CrossTable(x=Validation\_Label\_q5,y= Pred.Model.validation, prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1500   
##   
##   
## | Pred.Model.validation   
## Validation\_Label\_q5 | Fail | Success | Row Total |   
## --------------------|-----------|-----------|-----------|  
## Fail | 1351 | 5 | 1356 |   
## | 0.996 | 0.004 | 0.904 |   
## | 0.961 | 0.053 | |   
## | 0.901 | 0.003 | |   
## --------------------|-----------|-----------|-----------|  
## Success | 55 | 89 | 144 |   
## | 0.382 | 0.618 | 0.096 |   
## | 0.039 | 0.947 | |   
## | 0.037 | 0.059 | |   
## --------------------|-----------|-----------|-----------|  
## Column Total | 1406 | 94 | 1500 |   
## | 0.937 | 0.063 | |   
## --------------------|-----------|-----------|-----------|  
##   
##

confusionMatrix(Pred.Model.validation,Validation\_Label\_q5)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Fail Success  
## Fail 1351 55  
## Success 5 89  
##   
## Accuracy : 0.96   
## 95% CI : (0.9488, 0.9693)  
## No Information Rate : 0.904   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.7272   
##   
## Mcnemar's Test P-Value : 2.518e-10   
##   
## Sensitivity : 0.9963   
## Specificity : 0.6181   
## Pos Pred Value : 0.9609   
## Neg Pred Value : 0.9468   
## Prevalence : 0.9040   
## Detection Rate : 0.9007   
## Detection Prevalence : 0.9373   
## Balanced Accuracy : 0.8072   
##   
## 'Positive' Class : Fail   
##

#data Normalization on (Train+Validation) set and apply it to Test set (z-score scaling using “center” and “scale” as input method parameters)

# Use combined set to normalize  
norm.values.combined.q5 <- preProcess(TraVal\_Data\_q5, method=c("center", "scale"))  
Traval\_Norm\_q5<-predict(norm.values.combined.q5, TraVal\_Data\_q5)  
Test\_Norm\_q5<-predict(norm.values.combined.q5, Test\_Data\_q5)  
  
# Check the mean and variance of variables in the other sets  
summary(Traval\_Norm\_q5)

## Personal.Loan Age Experience Income   
## Length:4000 Min. :-1.95647 Min. :-2.020249 Min. :-1.4265   
## Class :character 1st Qu.:-0.90184 1st Qu.:-0.880141 1st Qu.:-0.7528   
## Mode :character Median :-0.02298 Median :-0.003135 Median :-0.2094   
## Mean : 0.00000 Mean : 0.000000 Mean : 0.0000   
## 3rd Qu.: 0.85587 3rd Qu.: 0.873870 3rd Qu.: 0.5295   
## Max. : 1.91050 Max. : 2.013978 Max. : 3.2680   
## Family CCAvg Mortgage Securities.Account  
## Min. :-1.2157 Min. :-1.1171 Min. :-0.5538 Min. :-0.3314   
## 1st Qu.:-1.2157 1st Qu.:-0.7107 1st Qu.:-0.5538 1st Qu.:-0.3314   
## Median :-0.3490 Median :-0.2463 Median :-0.5538 Median :-0.3314   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.5176 3rd Qu.: 0.3342 3rd Qu.: 0.4247 3rd Qu.:-0.3314   
## Max. : 1.3843 Max. : 4.6882 Max. : 5.6598 Max. : 3.0164   
## CD.Account Online CreditCard header   
## Min. :-0.2515 Min. :-1.2132 Min. :-0.6518 Mode:logical   
## 1st Qu.:-0.2515 1st Qu.:-1.2132 1st Qu.:-0.6518 TRUE:4000   
## Median :-0.2515 Median : 0.8241 Median :-0.6518   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.2515 3rd Qu.: 0.8241 3rd Qu.: 1.5337   
## Max. : 3.9753 Max. : 0.8241 Max. : 1.5337   
## Education\_1 Education\_2 Education\_3   
## Min. :-0.8552 Min. :-0.6239 Min. :-0.6503   
## 1st Qu.:-0.8552 1st Qu.:-0.6239 1st Qu.:-0.6503   
## Median :-0.8552 Median :-0.6239 Median :-0.6503   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 1.1690 3rd Qu.: 1.6024 3rd Qu.: 1.5374   
## Max. : 1.1690 Max. : 1.6024 Max. : 1.5374

var(Traval\_Norm\_q5[,2:14])

## Age Experience Income Family  
## Age 1.000000000 0.994253646 -0.050018488 -0.049450110  
## Experience 0.994253646 1.000000000 -0.040904814 -0.055890215  
## Income -0.050018488 -0.040904814 1.000000000 -0.161147080  
## Family -0.049450110 -0.055890215 -0.161147080 1.000000000  
## CCAvg -0.036166041 -0.033526194 0.635249961 -0.105757974  
## Mortgage -0.018602331 -0.017156876 0.210786796 -0.022030953  
## Securities.Account 0.003564431 0.002191085 0.007959293 0.017058277  
## CD.Account 0.013725876 0.015803804 0.181680490 0.006545439  
## Online 0.026510158 0.025638352 0.017414776 0.010000058  
## CreditCard 0.013882718 0.015306736 0.002947557 0.009245401  
## header 0.000000000 0.000000000 0.000000000 0.000000000  
## Education\_1 -0.023129805 0.001135665 0.213254711 -0.113903968  
## Education\_2 -0.012385455 -0.014553243 -0.125946231 0.131966935  
## CCAvg Mortgage Securities.Account CD.Account  
## Age -0.036166041 -0.0186023306 3.564431e-03 0.013725876  
## Experience -0.033526194 -0.0171568758 2.191085e-03 0.015803804  
## Income 0.635249961 0.2107867962 7.959293e-03 0.181680490  
## Family -0.105757974 -0.0220309531 1.705828e-02 0.006545439  
## CCAvg 1.000000000 0.1003192850 2.712105e-02 0.131721046  
## Mortgage 0.100319285 1.0000000000 -9.590367e-03 0.076894905  
## Securities.Account 0.027121050 -0.0095903666 1.000000e+00 0.295247357  
## CD.Account 0.131721046 0.0768949046 2.952474e-01 1.000000000  
## Online -0.006763999 0.0003941556 3.104068e-04 0.174999813  
## CreditCard -0.005203034 -0.0072685037 -2.032242e-02 0.277243675  
## header 0.000000000 0.0000000000 0.000000e+00 0.000000000  
## Education\_1 0.155402217 0.0447616297 4.558510e-03 -0.022582448  
## Education\_2 -0.092364390 -0.0363644652 3.913954e-05 0.007766372  
## Online CreditCard header Education\_1  
## Age 0.0265101582 0.013882718 0 -0.0231298050  
## Experience 0.0256383516 0.015306736 0 0.0011356654  
## Income 0.0174147758 0.002947557 0 0.2132547112  
## Family 0.0100000583 0.009245401 0 -0.1139039676  
## CCAvg -0.0067639991 -0.005203034 0 0.1554022171  
## Mortgage 0.0003941556 -0.007268504 0 0.0447616297  
## Securities.Account 0.0003104068 -0.020322417 0 0.0045585103  
## CD.Account 0.1749998126 0.277243675 0 -0.0225824477  
## Online 1.0000000000 0.001746397 0 0.0006238869  
## CreditCard 0.0017463969 1.000000000 0 0.0143347604  
## header 0.0000000000 0.000000000 0 0.0000000000  
## Education\_1 0.0006238869 0.014334760 0 1.0000000000  
## Education\_2 0.0277240947 -0.008928683 0 -0.5337272002  
## Education\_2  
## Age -1.238546e-02  
## Experience -1.455324e-02  
## Income -1.259462e-01  
## Family 1.319669e-01  
## CCAvg -9.236439e-02  
## Mortgage -3.636447e-02  
## Securities.Account 3.913954e-05  
## CD.Account 7.766372e-03  
## Online 2.772409e-02  
## CreditCard -8.928683e-03  
## header 0.000000e+00  
## Education\_1 -5.337272e-01  
## Education\_2 1.000000e+00

summary(Test\_Norm\_q5)

## Personal.Loan Age Experience Income   
## Length:1000 Min. :-1.95647 Min. :-2.02025 Min. :-1.4265   
## Class :character 1st Qu.:-0.90184 1st Qu.:-0.88014 1st Qu.:-0.7528   
## Mode :character Median : 0.06490 Median : 0.08457 Median :-0.2094   
## Mean : 0.03379 Mean : 0.03019 Mean : 0.0151   
## 3rd Qu.: 0.94376 3rd Qu.: 0.87387 3rd Qu.: 0.6165   
## Max. : 1.91050 Max. : 1.83858 Max. : 2.8551   
## Family CCAvg Mortgage Securities.Account  
## Min. :-1.21571 Min. :-1.11711 Min. :-0.553765 Min. :-0.33144   
## 1st Qu.:-1.21571 1st Qu.:-0.71074 1st Qu.:-0.553765 1st Qu.:-0.33144   
## Median :-0.34905 Median :-0.18826 Median :-0.553765 Median :-0.33144   
## Mean :-0.02752 Mean : 0.03964 Mean :-0.004597 Mean : 0.09039   
## 3rd Qu.: 0.51761 3rd Qu.: 0.39227 3rd Qu.: 0.463884 3rd Qu.:-0.33144   
## Max. : 1.38428 Max. : 4.68821 Max. : 5.141154 Max. : 3.01641   
## CD.Account Online CreditCard header   
## Min. :-0.25149 Min. :-1.21319 Min. :-0.65185 Mode:logical   
## 1st Qu.:-0.25149 1st Qu.:-1.21319 1st Qu.:-0.65185 TRUE:1000   
## Median :-0.25149 Median : 0.82407 Median :-0.65185   
## Mean : 0.01902 Mean : 0.01324 Mean :-0.04644   
## 3rd Qu.:-0.25149 3rd Qu.: 0.82407 3rd Qu.: 1.53372   
## Max. : 3.97527 Max. : 0.82407 Max. : 1.53372   
## Education\_1 Education\_2 Education\_3   
## Min. :-0.8552 Min. :-0.623918 Min. :-0.65029   
## 1st Qu.:-0.8552 1st Qu.:-0.623918 1st Qu.:-0.65029   
## Median :-0.8552 Median :-0.623918 Median :-0.65029   
## Mean :-0.0334 Mean : 0.003896 Mean : 0.03227   
## 3rd Qu.: 1.1690 3rd Qu.: 1.602373 3rd Qu.: 1.53739   
## Max. : 1.1690 Max. : 1.602373 Max. : 1.53739

var(Test\_Norm\_q5[,2:14])

## Age Experience Income Family  
## Age 1.07490136 1.06009145 -0.078947787 -0.03405533  
## Experience 1.06009145 1.05794444 -0.071136616 -0.03872465  
## Income -0.07894779 -0.07113662 1.005818383 -0.13889354  
## Family -0.03405533 -0.03872465 -0.138893537 0.94685521  
## CCAvg -0.12232810 -0.12244810 0.737857704 -0.12757900  
## Mortgage 0.01167761 0.01578723 0.186762447 -0.01318856  
## Securities.Account -0.01896606 -0.01730496 -0.046365390 0.03558999  
## CD.Account -0.01462981 -0.01125056 0.128268462 0.04494080  
## Online -0.03743862 -0.03304384 0.001211767 0.01176721  
## CreditCard -0.01575518 -0.01520741 -0.023128208 0.01940222  
## header 0.00000000 0.00000000 0.000000000 0.00000000  
## Education\_1 -0.04634928 -0.01835311 0.237059609 -0.13456025  
## Education\_2 -0.03254561 -0.03345194 -0.138768533 0.16489683  
## CCAvg Mortgage Securities.Account CD.Account  
## Age -0.122328100 0.011677615 -0.018966058 -0.0146298143  
## Experience -0.122448098 0.015787229 -0.017304956 -0.0112505566  
## Income 0.737857704 0.186762447 -0.046365390 0.1282684618  
## Family -0.127578997 -0.013188555 0.035589989 0.0449408045  
## CCAvg 1.146634243 0.153910062 -0.033019723 0.1701578502  
## Mortgage 0.153910062 0.953837990 0.011132869 0.1401878026  
## Securities.Account -0.033019723 0.011132869 1.235514699 0.4523645551  
## CD.Account 0.170157850 0.140187803 0.452364555 1.0712873623  
## Online 0.008333155 -0.031369662 0.062455680 0.1850803851  
## CreditCard -0.011518337 -0.006937508 0.008042051 0.2891760108  
## header 0.000000000 0.000000000 0.000000000 0.0000000000  
## Education\_1 0.175080688 0.033825965 0.019292396 0.0172658999  
## Education\_2 -0.089334287 -0.012880650 0.025873869 -0.0004521324  
## Online CreditCard header Education\_1 Education\_2  
## Age -0.037438621 -0.015755178 0 -0.04634928 -0.0325456077  
## Experience -0.033043839 -0.015207412 0 -0.01835311 -0.0334519411  
## Income 0.001211767 -0.023128208 0 0.23705961 -0.1387685327  
## Family 0.011767213 0.019402224 0 -0.13456025 0.1648968286  
## CCAvg 0.008333155 -0.011518337 0 0.17508069 -0.0893342874  
## Mortgage -0.031369662 -0.006937508 0 0.03382597 -0.0128806497  
## Securities.Account 0.062455680 0.008042051 0 0.01929240 0.0258738689  
## CD.Account 0.185080385 0.289176011 0 0.01726590 -0.0004521324  
## Online 0.995417327 0.014467479 0 0.01481115 -0.0080086758  
## CreditCard 0.014467479 0.957593341 0 0.01566798 -0.0249081553  
## header 0.000000000 0.000000000 0 0.00000000 0.0000000000  
## Education\_1 0.014811153 0.015667978 0 0.98914441 -0.5164735299  
## Education\_2 -0.008008676 -0.024908155 0 -0.51647353 1.0045514441

#modeling k-NN, Define predictors and labels for modeling on Train+Validation and test sets

TraVal\_Predictors\_q5<- TraVal\_Norm\_q5[,2:14]  
TraVal\_Label\_q5 <- TraVal\_Norm\_q5[,1]  
TraVal\_Label\_q5<-as.factor(TraVal\_Label\_q5)  
Test\_Predictors\_q5<-Test\_Norm\_q5[,2:14]  
Test\_Label\_q5<- Test\_Norm\_q5[,1]  
Test\_Label\_q5<-as.factor(Test\_Label\_q5)

#train a knn model

Pred.Model.test<-knn(TraVal\_Predictors\_q5, Test\_Predictors\_q5, cl=TraVal\_Label\_q5,k=3)

#confusion matrix for the test data

CrossTable(x=Test\_Label\_q5,y= Pred.Model.test, prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1000   
##   
##   
## | Pred.Model.test   
## Test\_Label\_q5 | Fail | Row Total |   
## --------------|-----------|-----------|  
## Fail | 904 | 904 |   
## | 0.904 | |   
## --------------|-----------|-----------|  
## Success | 96 | 96 |   
## | 0.096 | |   
## --------------|-----------|-----------|  
## Column Total | 1000 | 1000 |   
## --------------|-----------|-----------|  
##   
##

confusionMatrix(Pred.Model.test,Test\_Label\_q5)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Fail Success  
## Fail 904 96  
## Success 0 0  
##   
## Accuracy : 0.904   
## 95% CI : (0.884, 0.9215)  
## No Information Rate : 0.904   
## P-Value [Acc > NIR] : 0.5271   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 1.000   
## Specificity : 0.000   
## Pos Pred Value : 0.904   
## Neg Pred Value : NaN   
## Prevalence : 0.904   
## Detection Rate : 0.904   
## Detection Prevalence : 1.000   
## Balanced Accuracy : 0.500   
##   
## 'Positive' Class : Fail   
##

#Answer for Q5: The best k from Q2 was 3, and hyperparameter tuning on the new datasets (50% training, 30% validation, 20% test) also confirmed k=3 for the k-NN model.Accuracy based on Training and Validation sets: Training = 0.9728, Validation = 0.96. Accuracy based on (Training + Validation) and Test sets: Test = 0.904. The model performs best on the training set, since it was trained on that data. Validation accuracy is slightly lower, and test accuracy is the lowest because the model was tuned for the validation set, not the test set that has not been seen yet.