Assignment2 Nadukula Akanksha

2022-10-02

```r
library('caret')
```

```
## Warning: package 'caret' was built under R version 4.1.3
```

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

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'pillar'
```

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

```r
library('ISLR')
```

```
## Warning: package 'ISLR' was built under R version 4.1.3
```

```r
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
```

```r
library('class')

BankData <- read.csv("UniversalBank.csv" )

BankData$ID <- NULL
BankData$ZIP.Code <- NULL
summary(BankData)
```

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

```
#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 us
ing k = 1. Remember to transform categorical predictors with more than two categories into du
mmy variables first. Specify the success class as 1 (loan acceptance), and use the default cu
toff value of 0.5.

BankData$Personal.Loan =  as.factor(BankData$Personal.Loan)



Normalized_model <- preProcess(BankData[,-8],method = c("center", "scale"))
Bank_normalized <- predict(Normalized_model,BankData)
summary(Bank_normalized)
```

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

```
Train_index <- createDataPartition(BankData$Personal.Loan, p = 0.6, list = FALSE)
train.df = Bank_normalized[Train_index,]
validation.df = Bank_normalized[-Train_index,]


To_Predict = data.frame(Age = 40, Experience = 10, Income = 84, Family = 2,
                        CCAvg = 2, Education = 1, Mortgage = 0, Securities.Account =
                          0, CD.Account = 0, Online = 1, CreditCard = 1)
print(To_Predict)
```

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

```
To_Predict_Normalized <- predict(Normalized_model,To_Predict)

Prediction <- knn(train= train.df[,1:7,9:12],
                  test = To_Predict_Normalized[,1:7,9:12],
                  cl= train.df$Personal.Loan,
                  k=1)
print(Prediction)
```

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

```
#Question 2
#What is a choice of k that balances between overfitting and ignoring the predictor informati
on?
set.seed(123)
Bankcontrol <- trainControl(method= "repeatedcv", number = 3, repeats = 2)
searchGrid = expand.grid(k=1:10)

knn.model = train(Personal.Loan~., data = train.df, method = 'knn', tuneGrid = searchGrid,trC
ontrol = Bankcontrol)

knn.model
```

```
## k-Nearest Neighbors
##
## 3000 samples
##   11 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (3 fold, repeated 2 times)
## Summary of sample sizes: 2000, 2000, 2000, 2000, 2000, 2000, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    1  0.9525000  0.6935670
##    2  0.9465000  0.6570880
##    3  0.9525000  0.6743836
##    4  0.9505000  0.6578497
##    5  0.9525000  0.6680164
##    6  0.9513333  0.6625063
##    7  0.9500000  0.6445879
##    8  0.9481667  0.6283005
##    9  0.9476667  0.6222620
##   10  0.9455000  0.6012505
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 5.
```

```
#Question3

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

predictions <- predict(knn.model,validation.df)

confusionMatrix(predictions,validation.df$Personal.Loan)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1801   82
##          1    7  110
##
##                Accuracy : 0.9555
##                  95% CI : (0.9455, 0.9641)
##     No Information Rate : 0.904
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6894
##
##  Mcnemar's Test P-Value : 4.365e-15
##
##             Sensitivity : 0.9961
##             Specificity : 0.5729
##          Pos Pred Value : 0.9565
##          Neg Pred Value : 0.9402
##              Prevalence : 0.9040
##          Detection Rate : 0.9005
##    Detection Prevalence : 0.9415
##       Balanced Accuracy : 0.7845
##
##        'Positive' Class : 0
##
```

```
#Question4

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

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

```
#Question5
#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.
train_size = 0.5
Train_index = createDataPartition(BankData$Personal.Loan, p = 0.5, list = FALSE)
train.df = Bank_normalized[Train_index,]


test_size = 0.2
Test_index = createDataPartition(BankData$Personal.Loan, p = 0.2, list = FALSE)
Test.df = Bank_normalized[Test_index,]


valid_size = 0.3
Validation_index = createDataPartition(BankData$Personal.Loan, p = 0.3, list = FALSE)
validation.df = Bank_normalized[Validation_index,]



Testknn <- knn(train = train.df[,-8], test = Test.df[,-8], cl = train.df[,8], k =3)
Validationknn <- knn(train = train.df[,-8], test = validation.df[,-8], cl = train.df[,8], k =
3)
Trainknn <- knn(train = train.df[,-8], test = train.df[,-8], cl = train.df[,8], k =3)

confusionMatrix(Testknn, Test.df[,8])
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 901  28
##          1   3  68
##
##                Accuracy : 0.969
##                  95% CI : (0.9563, 0.9788)
##     No Information Rate : 0.904
##     P-Value [Acc > NIR] : 1.027e-15
##
##                   Kappa : 0.7979
##
##  Mcnemar's Test P-Value : 1.629e-05
##
##             Sensitivity : 0.9967
##             Specificity : 0.7083
##          Pos Pred Value : 0.9699
##          Neg Pred Value : 0.9577
##              Prevalence : 0.9040
##          Detection Rate : 0.9010
##    Detection Prevalence : 0.9290
##       Balanced Accuracy : 0.8525
##
##        'Positive' Class : 0
##
```

```
confusionMatrix(Trainknn, train.df[,8])
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2254   60
##          1    6  180
##
##                Accuracy : 0.9736
##                  95% CI : (0.9665, 0.9795)
##     No Information Rate : 0.904
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8309
##
##  Mcnemar's Test P-Value : 6.853e-11
##
##             Sensitivity : 0.9973
##             Specificity : 0.7500
##          Pos Pred Value : 0.9741
##          Neg Pred Value : 0.9677
##              Prevalence : 0.9040
##          Detection Rate : 0.9016
##    Detection Prevalence : 0.9256
##       Balanced Accuracy : 0.8737
##
##        'Positive' Class : 0
##
```

```
confusionMatrix(Validationknn, validation.df[,8])
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1349   41
##          1    7  103
##
##                Accuracy : 0.968
##                  95% CI : (0.9578, 0.9763)
##     No Information Rate : 0.904
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7939
##
##  Mcnemar's Test P-Value : 1.906e-06
##
##             Sensitivity : 0.9948
##             Specificity : 0.7153
##          Pos Pred Value : 0.9705
##          Neg Pred Value : 0.9364
##              Prevalence : 0.9040
##          Detection Rate : 0.8993
##    Detection Prevalence : 0.9267
##       Balanced Accuracy : 0.8551
##
##        'Positive' Class : 0
##
```