Assignment_2_KNN

SUPRIYA MATTAPELLY

2023-10-01

KSU ID : 811260002

Summary

Questions - Answers

1. How would this customer be classified? This new customer would be classified as 0, does not take the personal loan
2. The best K is 3

Problem Statement

Universal bank is a young bank growing rapidly in terms of overall customer acquisition. The majority of these customers are liability customers (depositors) with varying sizes of relationship with the bank. The customer base of asset customers (borrowers) is quite small, and the bank is interested in expanding this base rapidly in more loan business. In particular, it wants to explore ways of converting its liability customers to personal loan customers.

A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise smarter campaigns with better target marketing. The goal is to use k-NN to predict whether a new customer will accept a loan offer. This will serve as the basis for the design of a new campaign.

The file UniversalBank.csv contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

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

---

Data Import and Cleaning

First, load the required libraries

```
library(class)
library(caret)
```

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

```r
library(e1071)
```

Read the data.

```r
universal.df <- read.csv("UniversalBank.csv")
dim(universal.df)
```

```
## [1] 5000   14
```

```r
t(t(names(universal.df))) # The t function creates a transpose of the dataframe
```

```
##       [,1]
##  [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"
```

Drop ID and ZIP

```r
universal.df <- universal.df[,-c(1,5)]
```

Split Data into 60% training and 40% validation. There are many ways to do this. We will look at 2 different ways. Before we split, let us transform categorical variables into dummy variables

```r
# Only Education needs to be converted to factor
universal.df$Education <- as.factor(universal.df$Education)

# Now, convert Education to Dummy Variables

groups <- dummyVars(~., data = universal.df) # This creates the dummy groups
universal_m.df <- as.data.frame(predict(groups,universal.df))


set.seed(1)  # Important to ensure that we get the same sample if we rerun the code
train.index <- sample(row.names(universal_m.df), 0.6*dim(universal_m.df)[1])
valid.index <- setdiff(row.names(universal_m.df), train.index)
train.df <- universal_m.df[train.index,]
valid.df <- universal_m.df[valid.index,]
t(t(names(train.df)))
```

```
##       [,1]
##  [1,] "Age"
##  [2,] "Experience"
##  [3,] "Income"
##  [4,] "Family"
##  [5,] "CCAvg"
##  [6,] "Education.1"
##  [7,] "Education.2"
##  [8,] "Education.3"
##  [9,] "Mortgage"
## [10,] "Personal.Loan"
## [11,] "Securities.Account"
## [12,] "CD.Account"
## [13,] "Online"
## [14,] "CreditCard"
```

Now, let us normalize the data

```
train.norm.df <- train.df[,-10] # Note that Personal Income is the 10th variable
valid.norm.df <- valid.df[,-10]

norm.values <- preProcess(train.df[, -10], method=c("center", "scale"))
train.norm.df <- predict(norm.values, train.df[, -10])
valid.norm.df <- predict(norm.values, valid.df[, -10])
```

Questions

Consider the following customer:

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?

```
# We have converted all categorical variables to dummy variables
# Let's create a new sample
new_customer <- 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
)

# Normalize the new customer
new.cust.norm <- new_customer
new.cust.norm <- predict(norm.values, new.cust.norm)
```

Now, let us predict using knn

```
knn.pred1 <- class::knn(train = train.norm.df,
                        test = new.cust.norm,
                        cl = train.df$Personal.Loan, k = 1)
knn.pred1
```

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

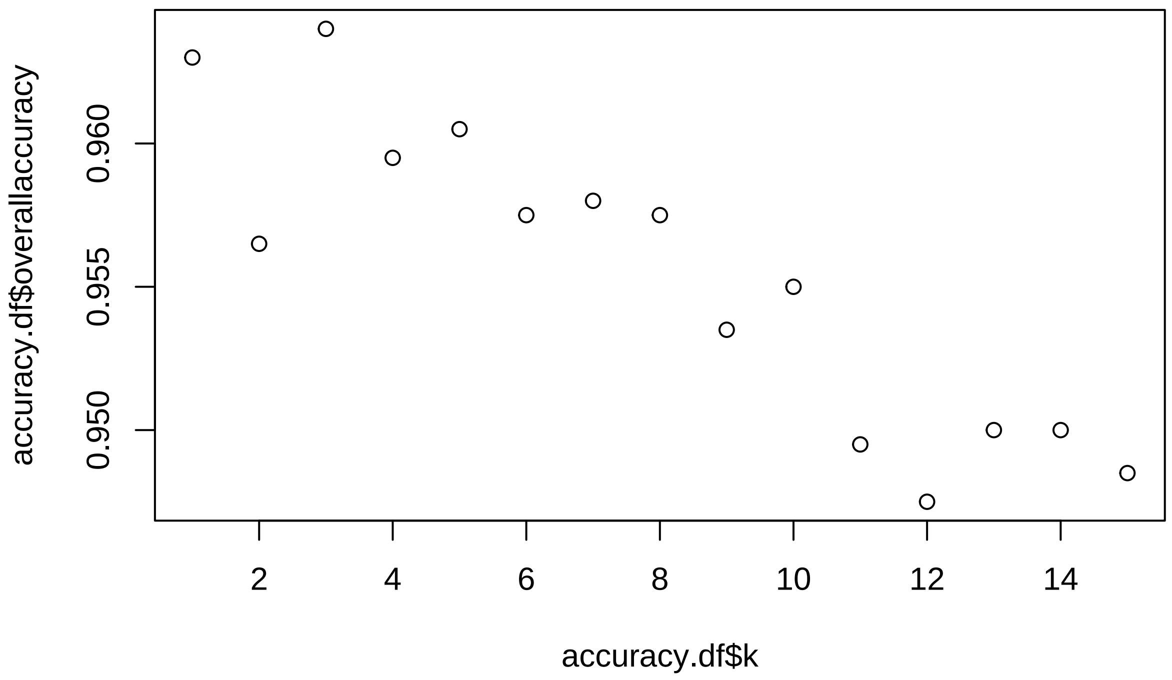
---

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

```
# Calculate the accuracy for each value of k
# Set the range of k values to consider

accuracy.df <- data.frame(k = seq(1, 15, 1), overallaccuracy = rep(0, 15))
for(i in 1:15) {
  knn.pred <- class::knn(train = train.norm.df,
                         test = valid.norm.df,
                         cl = train.df$Personal.Loan, k = i)
  accuracy.df[i, 2] <- confusionMatrix(knn.pred,
                as.factor(valid.df$Personal.Loan),positive = "1")$overall[1]

}

which(accuracy.df[,2] == max(accuracy.df[,2]))
```

```
## [1] 3
```

```
plot(accuracy.df$k,accuracy.df$overallaccuracy)
```



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

```
knn.valid <- knn(train = train.df[,-10],
             test = valid.df[,-10],
             cl = train.df[,10],
             k=3, prob=TRUE)



confusionMatrix(knn.valid,as.factor(valid.df[, 10]))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1725  131
##          1   70   74
##
##                Accuracy : 0.8995
##                  95% CI : (0.8855, 0.9123)
##     No Information Rate : 0.8975
##     P-Value [Acc > NIR] : 0.4017
##
##                   Kappa : 0.3709
##
```

```
##  Mcnemar's Test P-Value : 2.315e-05
##
##             Sensitivity : 0.9610
##             Specificity : 0.3610
##          Pos Pred Value : 0.9294
##          Neg Pred Value : 0.5139
##              Prevalence : 0.8975
##          Detection Rate : 0.8625
##    Detection Prevalence : 0.9280
##       Balanced Accuracy : 0.6610
##
##        'Positive' Class : 0
##
```

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.

```
customer.df= 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)

knn.pred2 <- knn(train = train.df[,-10],
             test = customer.df,
             cl = train.df[,10],
             k=3, prob=TRUE)
knn.pred2
```

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

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.

```
## Already we converted Education to Dummy Variables initially

set.seed(3) #setting to 3 as we had training, validation and testing sets

train.index <- sample(rownames(universal_m.df), 0.5*dim(universal_m.df)[1]) #training 50%
```

```r
valid.index <- sample(setdiff(rownames(universal_m.df),train.index),
                      0.3*dim(universal_m.df)[1]) #validation 30%


test.index = setdiff(rownames(universal_m.df), union(train.index, valid.index))




train.df<- universal_m.df[train.index, ]
valid.df <- universal_m.df[valid.index, ]
test.df <- universal_m.df[test.index, ]



#normalization of values


train.norm.df <- train.df[,-10] # Note that Personal Income is the 10th variable
valid.norm.df <- valid.df[,-10]
test.norm.df <- test.df[,-10]


norm.values <- preProcess(train.df[, -10], method=c("center", "scale"))
train.norm.df <- predict(norm.values, train.df[, -10])
valid.norm.df <- predict(norm.values, valid.df[, -10])
test.norm.df <- predict(norm.values, test.df[, -10])


#let us predict for Testing, Validation and Training sets using knn with k =3

knn.test.df <- knn(train = train.df[,-10],
                   test = test.df[,-10],
                   cl = train.df[,10],
                   k=3, prob=TRUE)                #knn of testing set



knn.valid.df <- knn(train = train.df[,-10],
                    test = valid.df[,-10],
                    cl = train.df[,10],
                    k=3, prob=TRUE)               #knn of validation set



knn.train.df <- knn(train = train.df[,-10],
                    test = train.df[,-10],
                    cl = train.df[,10],
                    k=3, prob=TRUE)               #knn of training set
```

```r
#confusion matrix of training set

confusionMatrix(knn.test.df, as.factor(test.df[,10]))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 865  68
##          1  34  33
##
##                Accuracy : 0.898
##                  95% CI : (0.8776, 0.9161)
##     No Information Rate : 0.899
##     P-Value [Acc > NIR] : 0.567890
##
##                   Kappa : 0.3397
##
##  Mcnemar's Test P-Value : 0.001085
##
##             Sensitivity : 0.9622
##             Specificity : 0.3267
##          Pos Pred Value : 0.9271
##          Neg Pred Value : 0.4925
##              Prevalence : 0.8990
##          Detection Rate : 0.8650
##    Detection Prevalence : 0.9330
##       Balanced Accuracy : 0.6445
##
##        'Positive' Class : 0
##
```

```r
#confusion matrix of validation set

confusionMatrix(knn.valid.df, as.factor(valid.df[,10]))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1312   97
##          1   40   51
##
##                Accuracy : 0.9087
##                  95% CI : (0.8929, 0.9228)
##     No Information Rate : 0.9013
##     P-Value [Acc > NIR] : 0.1821
##
##                   Kappa : 0.3802
```

```
##
##  Mcnemar's Test P-Value : 1.715e-06
##
##             Sensitivity : 0.9704
##             Specificity : 0.3446
##          Pos Pred Value : 0.9312
##          Neg Pred Value : 0.5604
##              Prevalence : 0.9013
##          Detection Rate : 0.8747
##    Detection Prevalence : 0.9393
##       Balanced Accuracy : 0.6575
##
##        'Positive' Class : 0
##
```

```
#confusion matrix of testing set

confusionMatrix(knn.train.df, as.factor(train.df[,10]))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2238   85
##          1   31  146
##
##                Accuracy : 0.9536
##                  95% CI : (0.9446, 0.9615)
##     No Information Rate : 0.9076
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6909
##
##  Mcnemar's Test P-Value : 8.614e-07
##
##             Sensitivity : 0.9863
##             Specificity : 0.6320
##          Pos Pred Value : 0.9634
##          Neg Pred Value : 0.8249
##              Prevalence : 0.9076
##          Detection Rate : 0.8952
##    Detection Prevalence : 0.9292
##       Balanced Accuracy : 0.8092
##
##        'Positive' Class : 0
##
```

Comparing the confusion matrix of test that of the training and validation sets.

1.test model have low accuracy compared to the validation and training sets which means predication are less accurate

2. test case have low kappa
3. sensitivity of the test case is also low compared to them which indicates the low ability of identifying the positive cases

4.Specificity is also low which indicates the low ability of identifying the negative cases

overall the test case confusion matrix has low in almost when compared to them, the differences may be due the factors like selection of data set, setting the parameters and method of approach