Final_Exam

Yash_Patel

2022-12-15

What we want to achieve with this analysis and which machine learning method we are going to use to predict?

Problem And Solution:

The bank's manager is concerned that more and more customers are abandoning their credit card services. They would really appreciate it if someone could foretell who was going to leave so they could go out of their way to offer better services and sway the customer's decision.

Therefore, in order to resolve the bank manager's dilemma, we will use the KNN classification machine learning algorithm to predict whether or not a given customer will discontinue using their credit card services. And using this, the bank manager can specifically target that customer in order to offer them better services and obtain feedback in to improve their service.

Loading the required packages

```
library (caret)
```

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

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
library(class)
```

From where I have collected the relevant real world data for the analysis and prediction?

I have used the data from the kaggle website. The link or reference of the data is as follows:https://www.kaggle.com/code/khyatigajera/creditcard-eda-knn-svm

Importing the dataset

```
BankChurnersALL <- read.csv("C:/Users/YASH/Downloads/Bankchurners - Sheet1.csv", header=TRUE)
summary(BankChurnersALL)
```

```
##  Attrition_Flag    Customer_Age       Gender       Dependent_count
##  Min.   :0.0000   Min.   :26.00   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:41.00   1st Qu.:0.0000   1st Qu.:1.000
##  Median :0.0000   Median :46.00   Median :0.0000   Median :2.000
##  Mean   :0.1572   Mean   :46.35   Mean   :0.4766   Mean   :2.338
##  3rd Qu.:0.0000   3rd Qu.:52.00   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :73.00   Max.   :1.0000   Max.   :5.000
##
##    Uneducated    High.School        College          Graduate
##  Min.   :0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.17   Mean   :0.2334   Mean   :0.1192   Mean   :0.3659
##  3rd Qu.:0.00   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##  Post.Graduate       Doctorate          Married           Single
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.06087   Mean   :0.05056   Mean   :0.5033   Mean   :0.4163
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##
##     Divorced       Less.than..40K    X.40K....60K     X.60K....80K
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.08036   Mean   :0.3943   Mean   :0.1994   Mean   :0.1558
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##  X.80K....120K       X.120K..       Category_Blue    Category_Silver
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :1.0000   Median :0.00000
```

```
##  Mean   :0.1698   Mean   :0.08078   Mean   :0.9318   Mean   :0.05522
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##
##  Category_Gold     Category_Platinum   Credit_Limit   Total_Trans_Amt
##  Min.   :0.00000   Min.   :0.000000   Min.   : 1438   Min.   :  510
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.: 2498   1st Qu.: 2089
##  Median :0.00000   Median :0.000000   Median : 4287   Median : 3831
##  Mean   :0.01144   Mean   :0.001553   Mean   : 8493   Mean   : 4394
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:10729   3rd Qu.: 4740
##  Max.   :1.00000   Max.   :1.000000   Max.   :34516   Max.   :17995
##
##     X               X.1                X.2
##  Mode:logical   Length:7081        Min.   :0.0
##  NA's:7081      Class :character   1st Qu.:0.0
##                 Mode  :character   Median :0.5
##                                    Mean   :0.5
##                                    3rd Qu.:1.0
##                                    Max.   :1.0
##                                    NA's   :7077
```

What did I do to improve the quality and effectiveness of the data?

Here, I used spreadsheets to convert the categorical data to numerical data in order to make the best prediction. In order to improve the data's accuracy and usefulness, I also used R to remove the columns that were significantly less important for predicting the churn rate.

Selecting the necessary column and droping the one's which are less useful for the analysis

```
BankChurners <- select(BankChurnersALL,-(25:27))
#Here we use the summary function to get the overview of the data.
summary(BankChurners)
```

```
##  Attrition_Flag    Customer_Age       Gender       Dependent_count
##  Min.   :0.0000   Min.   :26.00   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:41.00   1st Qu.:0.0000   1st Qu.:1.000
##  Median :0.0000   Median :46.00   Median :0.0000   Median :2.000
##  Mean   :0.1572   Mean   :46.35   Mean   :0.4766   Mean   :2.338
##  3rd Qu.:0.0000   3rd Qu.:52.00   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :73.00   Max.   :1.0000   Max.   :5.000
##    Uneducated    High.School        College          Graduate
##  Min.   :0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.17   Mean   :0.2334   Mean   :0.1192   Mean   :0.3659
##  3rd Qu.:0.00   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  Post.Graduate       Doctorate          Married           Single
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
```

```
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.06087   Mean   :0.05056   Mean   :0.5033   Mean   :0.4163
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##     Divorced       Less.than..40K    X.40K....60K     X.60K....80K
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.08036   Mean   :0.3943   Mean   :0.1994   Mean   :0.1558
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  X.80K....120K       X.120K..       Category_Blue    Category_Silver
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :1.0000   Median :0.00000
##  Mean   :0.1698   Mean   :0.08078   Mean   :0.9318   Mean   :0.05522
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##  Category_Gold     Category_Platinum   Credit_Limit   Total_Trans_Amt
##  Min.   :0.00000   Min.   :0.000000   Min.   : 1438   Min.   :  510
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.: 2498   1st Qu.: 2089
##  Median :0.00000   Median :0.000000   Median : 4287   Median : 3831
##  Mean   :0.01144   Mean   :0.001553   Mean   : 8493   Mean   : 4394
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:10729   3rd Qu.: 4740
##  Max.   :1.00000   Max.   :1.000000   Max.   :34516   Max.   :17995
```

How did I prepare the data for the model?

In order to improve the performance of the model, I first normalised the data to get all of the variable values in a single range. I then partitioned the data into two sets, using 70% of it as a training set and the remaining 30% as a testing set. finally used the Knn function.

Normalising the data

```
normalise <- function(x) {return((x-min(x))/(max(x)-min(x)))}
BankChurners.n <- as.data.frame(lapply(BankChurners, normalise))
summary(BankChurners.n)
```

```
##  Attrition_Flag    Customer_Age        Gender       Dependent_count
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.3191   1st Qu.:0.0000   1st Qu.:0.2000
##  Median :0.0000   Median :0.4255   Median :0.0000   Median :0.4000
##  Mean   :0.1572   Mean   :0.4329   Mean   :0.4766   Mean   :0.4676
##  3rd Qu.:0.0000   3rd Qu.:0.5532   3rd Qu.:1.0000   3rd Qu.:0.6000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##    Uneducated    High.School        College          Graduate
##  Min.   :0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00   Median :0.0000   Median :0.0000   Median :0.0000
```

```
##  Mean   :0.17   Mean   :0.2334   Mean   :0.1192   Mean   :0.3659
##  3rd Qu.:0.00   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  Post.Graduate       Doctorate          Married           Single
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.06087   Mean   :0.05056   Mean   :0.5033   Mean   :0.4163
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##     Divorced       Less.than..40K    X.40K....60K     X.60K....80K
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.08036   Mean   :0.3943   Mean   :0.1994   Mean   :0.1558
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  X.80K....120K       X.120K..       Category_Blue    Category_Silver
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :1.0000   Median :0.00000
##  Mean   :0.1698   Mean   :0.08078   Mean   :0.9318   Mean   :0.05522
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##  Category_Gold     Category_Platinum   Credit_Limit     Total_Trans_Amt
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.03204   1st Qu.:0.09031
##  Median :0.00000   Median :0.000000   Median :0.08612   Median :0.18993
##  Mean   :0.01144   Mean   :0.001553   Mean   :0.21327   Mean   :0.22215
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.28087   3rd Qu.:0.24192
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.00000   Max.   :1.00000
```

Selecting 70 % data through random sampling.

```
set.seed(10923)
data.1 <- sample(1:nrow(BankChurners.n),size = nrow(BankChurners.n) * 0.7,
                 replace = FALSE )
# 70 % training data
train.bc <- BankChurners[ data.1, ]
# 30 % test data
test.bc <- BankChurners[ -data.1, ]
```

Now creating seperate dataframe for ' atrrition flag' feature which is our target.

```
train.bc.1 <- BankChurners[data.1, 1]
test.bc.1 <- BankChurners[-data.1 , 1 ]
```

knn

```
NROW ( train.bc.1 )
```

```
## [1] 4956
```

```
#sqrt(4956)
#~70

knn.70 <- knn(train = train.bc, test = test.bc, cl = train.bc.1, k = 70 )
knn.71 <- knn(train = train.bc, test = test.bc, cl = train.bc.1, k = 71 )
```

How I have evaluated the model?

To evaluate the model, I checked the accuracy and used the confusion matrix to assess the model's performance. At last, to get the best K value, I initiated the loop that will show the K value with the highest accuracy.

Let's calculate the proportion of correct classification for k = 70, 71

```
ACC.70 <- 100 * sum ( test.bc.1 == knn.70 ) / NROW ( test.bc.1 )
ACC.71 <- 100 * sum ( test.bc.1 == knn.71 ) / NROW ( test.bc.1 )
ACC.70#85.83529
```

```
## [1] 85.83529
```

```
ACC.71#85.69412
```

```
## [1] 85.88235
```

Checking prediction

```
table (knn.70, test.bc.1)
```

```
##       test.bc.1
## knn.70    0    1
##      0 1753  270
##      1   31   71
```

```
table (knn.71, test.bc.1)
```

```
##       test.bc.1
## knn.71    0    1
##      0 1753  269
##      1   31   72
```

confusionMatrix

```
confusionMatrix (table(knn.70, test.bc.1))
```

```
## Confusion Matrix and Statistics
##
##       test.bc.1
## knn.70    0    1
##      0 1753  270
##      1   31   71
##
##                Accuracy : 0.8584
##                  95% CI : (0.8428, 0.8729)
##     No Information Rate : 0.8395
##     P-Value [Acc > NIR] : 0.008987
##
##                   Kappa : 0.2663
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9826
##             Specificity : 0.2082
##          Pos Pred Value : 0.8665
##          Neg Pred Value : 0.6961
##              Prevalence : 0.8395
##          Detection Rate : 0.8249
##    Detection Prevalence : 0.9520
##       Balanced Accuracy : 0.5954
##
##        'Positive' Class : 0
##
```

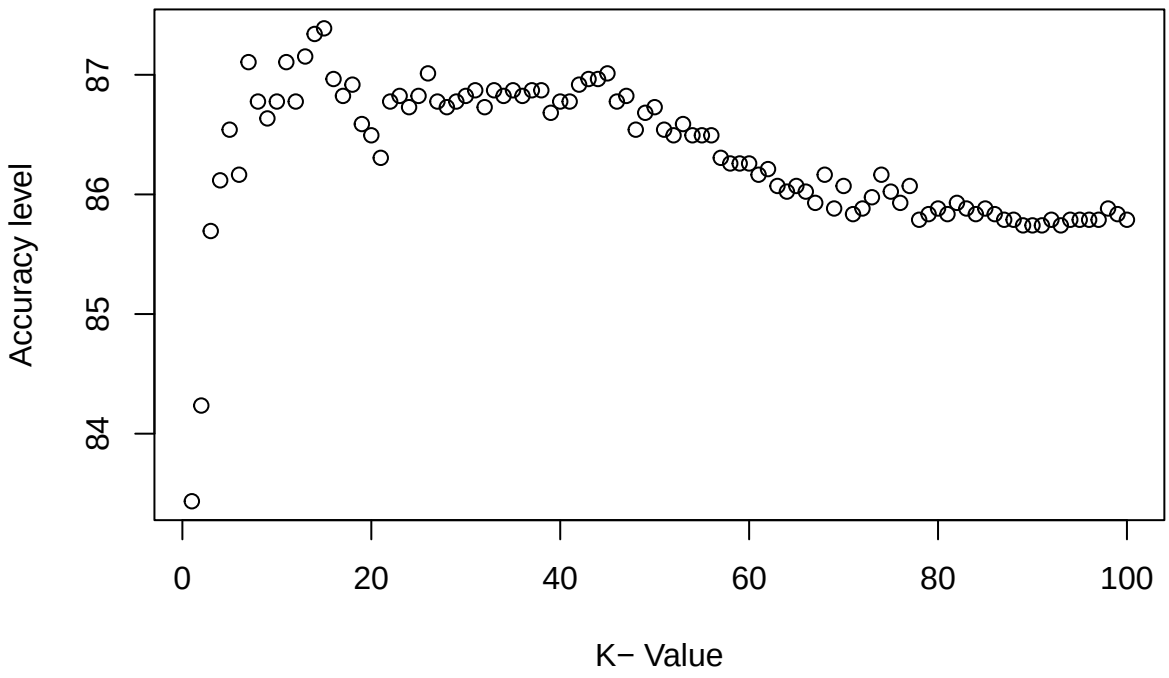
finding the best K value

```
# declaration to initiate for loop
i= 1
k.optm = 1
for ( i in 1:100 ) {
  knn.mod <- knn( train=train.bc, test=test.bc, cl=train.bc.1, k = i )
  k.optm [ i ] <- 100 * sum (test.bc.1 == knn.mod)/NROW(test.bc.1)
  k = i
  cat (k,' = ',k.optm [ i ],' \ n ') # to print % accuracy
}
```

```
## 1  =  83.43529   n 2  =  84.23529   n 3  =  85.69412   n 4  =  86.11765   n 5  =  86.54118   n 6  =
```

```
# ploting % accuracy  to k - value
plot( k.optm, type = "point", xlab="K- Value",ylab="Accuracy level")
```

```
## Warning in plot.xy(xy, type, ...): plot type 'point' will be truncated to first
## character
```



Here we got to know that when K value is 15 we get the highest accuracy #87.38824

```
knn.15 <- knn(train = train.bc, test = test.bc, cl = train.bc.1, k = 15 )
confusionMatrix (table(knn.15, test.bc.1))
```

```
## Confusion Matrix and Statistics
##
##       test.bc.1
## knn.15    0    1
##      0 1737  221
##      1   47  120
##
##                Accuracy : 0.8739
##                  95% CI : (0.859, 0.8877)
##     No Information Rate : 0.8395
##     P-Value [Acc > NIR] : 5.071e-06
##
##                   Kappa : 0.4102
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9737
##             Specificity : 0.3519
##          Pos Pred Value : 0.8871
##          Neg Pred Value : 0.7186
##              Prevalence : 0.8395
```

```
##          Detection Rate : 0.8174
##    Detection Prevalence : 0.9214
##       Balanced Accuracy : 0.6628
##
##        'Positive' Class : 0
##
```