Anna's World is setting up their business so they are not so reliant on selling gas. One way is to increase the sales of profitable products. I will use a KNN algorithm to address the business problem of predicting when a transaction will be a sale of a high gross profit margin.

Brian Estvander 17 MAY 2024

KNN ANALYSIS USING R

```
# Load packages----
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
## Warning: package 'tibble' was built under R version 4.3.3
## Warning: package 'tidyr' was built under R version 4.3.3
## Warning: package 'readr' was built under R version 4.3.3
## Warning: package 'purrr' was built under R version 4.3.3
## Warning: package 'dplyr' was built under R version 4.3.3
## Warning: package 'stringr' was built under R version 4.3.3
## Warning: package 'forcats' was built under R version 4.3.3
## Warning: package 'lubridate' was built under R version 4.3.3
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# install.packages('e1071')----
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.3.3
```

```
# Build Confusion Matrix Function----
my_confusion_matrix <- function(cf_table) {
  true_positive <- cf_table[4]
  true_negative <- cf_table[1]
  false_positive <- cf_table[2]
  false_negative <- cf_table[3]
  accuracy <- (true_positive + true_negative) / (true_positive + true_negative + false_positive + false_
  sensitivity_recall <- true_positive / (true_positive + false_negative)
  specificity_selectivity <- true_negative / (true_negative + false_positive)
  precision <- true_positive / (true_positive + false_positive)
  neg_pred_value <- true_negative/(true_negative + false_negative)
  print(cf_table)
  my_list <- list(sprintf("%1.0f = True Positive (TP), Hit", true_positive),
                  sprintf("%1.0f = True Negative (TN), Rejection", true_negative),
                  sprintf("%1.0f = False Positive (FP), Type 1 Error", false_positive),
                  sprintf("%1.0f = False Negative (FN), Type 2 Error", false_negative),
```

```
                  sprintf("%1.4f = Accuracy (TP+TN/(TP+TN+FP+FN))", accuracy),
                  sprintf("%1.4f = Sensitivity, Recall, Hit Rate, True Positive Rate (How many positives
                  sprintf("%1.4f = Specificity, Selectivity, True Negative Rate (How many negatives did
                  sprintf("%1.4f = Precision, Positive Predictive Value (How good are the model's positi
                  sprintf("%1.4f = Negative Predictive Value (How good are the model's negative predicti
  )
  return(my_list)
}
```

```
# Read in data----
knn_input <- read_rds('knn_input.rds')
str(knn_input)
```

```
## 'data.frame':    20000 obs. of  38 variables:
##  $ high_gpm                   : Factor w/ 2 levels "low","high": 2 1 1 2 2 2 1 1 1 1 ...
##  $ revenue                    : num  0.99 27.25 2.49 2.26 3 ...
##  $ quarter.1                  : int  1 0 1 1 0 0 0 0 0 0 ...
##  $ quarter.2                  : int  0 1 0 0 1 1 1 0 0 0 ...
##  $ quarter.3                  : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ quarter.4                  : int  0 0 0 0 0 0 0 1 0 1 ...
##  $ income                     : int  65268 58854 61115 45874 43547 41827 45874 66250 70963 61115 ...
##  $ bachelors_degree           : int  118 68 264 2488 502 169 2488 22 1959 264 ...
##  $ population                 : int  1003 734 2048 30327 10615 2379 30327 567 16052 2048 ...
##  $ state_province.Alabama     : int  0 0 0 0 1 0 0 0 1 0 ...
##  $ state_province.Arkansas    : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ state_province.Colorado    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ state_province.Iowa        : int  1 0 0 0 0 0 0 1 0 0 ...
##  $ state_province.Minnesota   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ state_province.Missouri    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ state_province.Nebraska    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ state_province.Oklahoma    : int  0 0 1 0 0 1 0 0 0 1 ...
##  $ state_province.South Dakota: int  0 1 0 0 0 0 0 0 0 0 ...
##  $ state_province.Wyoming     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ num_trans                  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ basket.no                  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ basket.yes                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ refill.no                  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ refill.yes                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ area.alcohol               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ area.cooler                : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ area.dispensed             : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ area.fresh                 : int  1 0 0 1 1 0 0 0 0 0 ...
##  $ area.fuel                  : int  0 1 0 0 0 0 1 1 0 1 ...
##  $ area.grocery               : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ area.lottery               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ area.miscellaneous         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ area.nongrocery            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ area.snacks                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ area.tobacco               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ items_sold                 : num  1 1 1 1 3 1 1 1 1 1 ...
##  $ loyalty2.not loyal         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ loyalty2.loyal             : int  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "dummies")=List of 6
##   ..$ quarter       : int [1:4] 3 4 5 6
```

```
##   ..$ state_province: int [1:10] 10 11 12 13 14 15 16 17 18 19
##   ..$ basket        : int [1:2] 21 22
##   ..$ refill        : int [1:2] 23 24
##   ..$ area          : int [1:11] 25 26 27 28 29 30 31 32 33 34 ...
##   ..$ loyalty2      : int [1:2] 37 38
```

```
slice_sample(knn_input, n=10)
```

```
##         high_gpm revenue quarter.1 quarter.2 quarter.3 quarter.4 income
## 459479      high    1.59         0         1         0         0  57600
## 1264513      low    5.00         1         0         0         0  34152
## 1514848      low   31.17         0         0         1         0  94875
## 810004       low    2.00         1         0         0         0  45651
## 313028      high    1.49         0         0         1         0  61492
## 2472         low    6.47         0         0         1         0  46818
## 93908       high    3.99         0         0         0         1  65625
## 425652       low   33.37         0         0         1         0  45874
## 457917      high    1.00         0         0         0         1  51042
## 588355       low    2.75         0         0         0         1  59500
##         bachelors_degree population state_province.Alabama
## 459479              3483      24180                      0
## 1264513               32       1904                      0
## 1514848             8366      34092                      0
## 810004               930       8350                      0
## 313028               137       1840                      0
## 2472                 261       3158                      0
## 93908                207       2270                      0
## 425652              2488      30327                      0
## 457917                47        329                      0
## 588355               876       8410                      0
##         state_province.Arkansas state_province.Colorado state_province.Iowa
## 459479                        0                       0                   0
## 1264513                       0                       0                   0
## 1514848                       0                       1                   0
## 810004                        0                       0                   1
## 313028                        0                       0                   1
## 2472                          0                       0                   0
## 93908                         0                       0                   1
## 425652                        1                       0                   0
## 457917                        0                       0                   0
## 588355                        1                       0                   0
##         state_province.Minnesota state_province.Missouri
## 459479                         0                      1
## 1264513                        0                      1
## 1514848                        0                      0
## 810004                         0                      0
## 313028                         0                      0
## 2472                           0                      0
## 93908                          0                      0
## 425652                         0                      0
## 457917                         0                      0
## 588355                         0                      0
##         state_province.Nebraska state_province.Oklahoma
## 459479                        0                      0
## 1264513                       0                      0
```

```
## 1514848                       0                       0
## 810004                        0                       0
## 313028                        0                       0
## 2472                          0                       1
## 93908                         0                       0
## 425652                        0                       0
## 457917                        1                       0
## 588355                        0                       0
##         state_province.South Dakota state_province.Wyoming num_trans basket.no
## 459479                            0                       0         1         1
## 1264513                           0                       0         1         1
## 1514848                           0                       0         1         1
## 810004                            0                       0         1         1
## 313028                            0                       0         1         1
## 2472                              0                       0         1         1
## 93908                             0                       0         1         1
## 425652                            0                       0         1         1
## 457917                            0                       0         1         1
## 588355                            0                       0         1         1
##         basket.yes refill.no refill.yes area.alcohol area.cooler area.dispensed
## 459479           0         1          0            0           0               1
## 1264513          0         1          0            0           0               0
## 1514848          0         1          0            0           0               0
## 810004           0         1          0            0           1               0
## 313028           0         1          0            0           1               0
## 2472             0         1          0            0           0               0
## 93908            0         1          0            0           0               0
## 425652           0         1          0            0           0               0
## 457917           0         1          0            0           0               1
## 588355           0         1          0            0           1               0
##         area.fresh area.fuel area.grocery area.lottery area.miscellaneous
## 459479           0         0            0            0                  0
## 1264513          0         1            0            0                  0
## 1514848          0         1            0            0                  0
## 810004           0         0            0            0                  0
## 313028           0         0            0            0                  0
## 2472             0         0            0            0                  0
## 93908            1         0            0            0                  0
## 425652           0         1            0            0                  0
## 457917           0         0            0            0                  0
## 588355           0         0            0            0                  0
##         area.nongrocery area.snacks area.tobacco items_sold loyalty2.not loyal
## 459479                0           0            0          1                  0
## 1264513               0           0            0          1                  1
## 1514848               0           0            0          1                  1
## 810004                0           0            0          1                  1
## 313028                0           0            0          1                  1
## 2472                  0           0            1          1                  0
## 93908                 0           0            0          1                  0
## 425652                0           0            0          1                  0
## 457917                0           0            0          1                  0
## 588355                0           0            0          1                  1
##         loyalty2.loyal
## 459479               1
```

```
## 1264513              0
## 1514848              0
## 810004               0
## 313028               0
## 2472                 1
## 93908                1
## 425652               1
## 457917               1
## 588355               0
```

```
# Explore the target feature

freq <- table(knn_input$high_gpm)
freq[2]/(freq[1]+freq[2])
```

```
##    high
## 0.43695
```

```
contrasts(knn_input$high_gpm)
```

```
##      high
## low     0
## high    1
```

```
# Partition the data

library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

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

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(77)
partition <- caret::createDataPartition(y=knn_input$high_gpm, p=.75, list=FALSE)
data_train <- knn_input[partition, ]
data_test <- knn_input[-partition, ]
```

```
# Separate the target variable

X_train <- data_train %>% select(-high_gpm)
X_test <-  data_test %>% select(-high_gpm)
y_train <- data_train$high_gpm
y_test <- data_test$high_gpm
```

```
# z-score standardization

X_train <- scale(X_train)
X_test <- scale(X_test)
```

```
# Double check sizes

nrow(X_train)/(nrow(X_test)+nrow(X_train))
```

```
## [1] 0.75005
```

```
dim(X_train)
```

```
## [1] 15001    37
```

```
length(y_train)
```

```
## [1] 15001
```

```
# I took the square root of the total number of rows to get my k.
# Run the model

library(class)
knn1 = class::knn(train=X_train, test=X_test, cl=y_train, k=141)
```

```
# Run Confusion matrix to check accuracy
# Prediction on left and truth on top

table2 <- table(knn1, y_test)
my_confusion_matrix(table2)
```

```
##       y_test
## knn1    low high
##   low  2448  484
##   high  367 1700
```

```
## [[1]]
## [1] "1700 = True Positive (TP), Hit"
##
## [[2]]
## [1] "2448 = True Negative (TN), Rejection"
##
## [[3]]
## [1] "367 = False Positive (FP), Type 1 Error"
##
## [[4]]
## [1] "484 = False Negative (FN), Type 2 Error"
##
## [[5]]
## [1] "0.8298 = Accuracy (TP+TN/(TP+TN+FP+FN))"
##
## [[6]]
## [1] "0.7784 = Sensitivity, Recall, Hit Rate, True Positive Rate (How many positives did the model ge
##
## [[7]]
## [1] "0.8696 = Specificity, Selectivity, True Negative Rate (How many negatives did the model get rig
##
## [[8]]
## [1] "0.8224 = Precision, Positive Predictive Value (How good are the model's positive predictions? T
##
## [[9]]
## [1] "0.8349 = Negative Predictive Value (How good are the model's negative predictions? TN/(TN+FN)"
```

```
# Pre-programmed confusion matrix

caret::confusionMatrix(knn1, y_test, positive='high')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  low high
##       low  2448  484
##       high  367 1700
##
##                Accuracy : 0.8298
##                  95% CI : (0.8191, 0.8401)
##     No Information Rate : 0.5631
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6519
##
##  Mcnemar's Test P-Value : 6.996e-05
##
##             Sensitivity : 0.7784
##             Specificity : 0.8696
##          Pos Pred Value : 0.8224
##          Neg Pred Value : 0.8349
##              Prevalence : 0.4369
##          Detection Rate : 0.3401
##    Detection Prevalence : 0.4135
##       Balanced Accuracy : 0.8240
##
##        'Positive' Class : high
##
```

```
# Evaluate the data

data_test$prediction <- knn1
data_test <- data_test %>% mutate(correct = high_gpm==prediction)
slice_sample(data_test, n=20)
```

```
##         high_gpm revenue quarter.1 quarter.2 quarter.3 quarter.4 income
## 209931      high    1.79         0         0         1         0  34210
## 1403545      low   26.91         0         0         0         1  29091
## 1226812      low   10.00         1         0         0         0  42772
## 428400       low    5.00         1         0         0         0  66250
## 417907       low   21.00         1         0         0         0  76071
## 38464       high    0.99         0         1         0         0  52500
## 1612610     high    1.00         0         0         0         1  29091
## 1224527      low   45.25         0         0         0         1  70557
## 879815      high    1.98         0         0         1         0  66160
## 318760       low    7.03         0         1         0         0  37130
## 1242817      low   26.90         0         0         0         1  50819
## 382899      high    1.69         0         0         1         0  66250
## 118951       low   13.98         1         0         0         0  41827
## 855741       low    4.75         0         1         0         0  70963
## 1551687     high    2.69         0         0         1         0  66250
## 1201431      low   20.00         0         0         0         1  51461
## 471531      high    7.56         0         0         1         0  63673
## 15110        low    2.13         0         0         0         1  53026
## 1555083     high    1.30         1         0         0         0 132230
## 417764       low   44.64         0         0         1         0  73148
```

```
##         bachelors_degree population state_province.Alabama
## 209931              1632      25999                       1
## 1403545                3        181                       0
## 1226812              160       1876                       0
## 428400                22        567                       0
## 417907                93        493                       0
## 38464                 95       1192                       0
## 1612610                3        181                       0
## 1224527             4048      51863                       0
## 879815             10178      57364                       0
## 318760              1238      15153                       0
## 1242817              958       8778                       0
## 382899                22        567                       0
## 118951               169       2379                       0
## 855741              1959      16052                       1
## 1551687               22        567                       0
## 1201431             2922      52172                       0
## 471531              1172       9046                       0
## 15110                 19        557                       0
## 1555083              974       3294                       0
## 417764               139        852                       0
##         state_province.Arkansas state_province.Colorado state_province.Iowa
## 209931                        0                       0                    0
## 1403545                       0                       0                    0
## 1226812                       0                       0                    0
## 428400                        0                       0                    1
## 417907                        0                       0                    0
## 38464                         0                       0                    0
## 1612610                       0                       0                    0
## 1224527                       0                       1                    0
## 879815                        0                       0                    0
## 318760                        0                       1                    0
## 1242817                       0                       0                    0
## 382899                        0                       0                    1
## 118951                        0                       0                    0
## 855741                        0                       0                    0
## 1551687                       0                       0                    1
## 1201431                       0                       1                    0
## 471531                        0                       0                    1
## 15110                         0                       0                    1
## 1555083                       1                       0                    0
## 417764                        0                       0                    1
##         state_province.Minnesota state_province.Missouri
## 209931                         0                      0
## 1403545                        0                      0
## 1226812                        0                      0
## 428400                         0                      0
## 417907                         0                      0
## 38464                          0                      0
## 1612610                        0                      0
## 1224527                        0                      0
## 879815                         0                      1
## 318760                         0                      0
## 1242817                        1                      0
```

```
## 382899                          0                         0
## 118951                          0                         0
## 855741                          0                         0
## 1551687                         0                         0
## 1201431                         0                         0
## 471531                          0                         0
## 15110                           0                         0
## 1555083                         0                         0
## 417764                          0                         0
##         state_province.Nebraska state_province.Oklahoma
## 209931                        0                       0
## 1403545                       0                       1
## 1226812                       0                       1
## 428400                        0                       0
## 417907                        1                       0
## 38464                         1                       0
## 1612610                       0                       1
## 1224527                       0                       0
## 879815                        0                       0
## 318760                        0                       0
## 1242817                       0                       0
## 382899                        0                       0
## 118951                        0                       1
## 855741                        0                       0
## 1551687                       0                       0
## 1201431                       0                       0
## 471531                        0                       0
## 15110                         0                       0
## 1555083                       0                       0
## 417764                        0                       0
##         state_province.South Dakota state_province.Wyoming num_trans basket.no
## 209931                            0                       0         1         1
## 1403545                           0                       0         1         1
## 1226812                           0                       0         1         1
## 428400                            0                       0         1         1
## 417907                            0                       0         1         1
## 38464                             0                       0         1         1
## 1612610                           0                       0         1         1
## 1224527                           0                       0         1         1
## 879815                            0                       0         1         1
## 318760                            0                       0         1         1
## 1242817                           0                       0         1         1
## 382899                            0                       0         1         1
## 118951                            0                       0         1         1
## 855741                            0                       0         1         1
## 1551687                           0                       0         1         1
## 1201431                           0                       0         1         1
## 471531                            0                       0         1         1
## 15110                             0                       0         1         1
## 1555083                           0                       0         1         1
## 417764                            0                       0         1         1
##         basket.yes refill.no refill.yes area.alcohol area.cooler area.dispensed
## 209931           0         1          0            0           0              0
## 1403545          0         1          0            0           0              0
```

```
## 1226812          0         1          0            0          0              0
## 428400           0         1          0            0          0              0
## 417907           0         1          0            0          0              0
## 38464            0         1          0            0          1              0
## 1612610          0         1          0            0          0              1
## 1224527          0         1          0            0          0              0
## 879815           0         1          0            0          0              0
## 318760           0         1          0            0          0              0
## 1242817          0         1          0            0          0              0
## 382899           0         1          0            0          0              1
## 118951           0         1          0            0          0              0
## 855741           0         1          0            0          0              0
## 1551687          0         1          0            0          0              0
## 1201431          0         1          0            0          0              0
## 471531           0         1          0            0          1              0
## 15110            0         1          0            0          1              0
## 1555083          0         1          0            0          0              0
## 417764           0         1          0            0          0              0
##         area.fresh area.fuel area.grocery area.lottery area.miscellaneous
## 209931           1         0            0           0                   0
## 1403545          0         1            0           0                   0
## 1226812          0         1            0           0                   0
## 428400           0         1            0           0                   0
## 417907           0         1            0           0                   0
## 38464            0         0            0           0                   0
## 1612610          0         0            0           0                   0
## 1224527          0         1            0           0                   0
## 879815           0         0            0           0                   0
## 318760           0         0            0           0                   0
## 1242817          0         1            0           0                   0
## 382899           0         0            0           0                   0
## 118951           0         0            0           0                   0
## 855741           0         0            0           0                   0
## 1551687          1         0            0           0                   0
## 1201431          0         1            0           0                   0
## 471531           0         0            0           0                   0
## 15110            0         0            0           0                   0
## 1555083          1         0            0           0                   0
## 417764           0         1            0           0                   0
##         area.nongrocery area.snacks area.tobacco items_sold loyalty2.not loyal
## 209931                0           0            0          1                   1
## 1403545               0           0            0          1                   1
## 1226812               0           0            0          1                   1
## 428400                0           0            0          1                   1
## 417907                0           0            0          1                   0
## 38464                 0           0            0          1                   0
## 1612610               0           0            0          1                   1
## 1224527               0           0            0          1                   1
## 879815                0           1            0          2                   1
## 318760                0           0            1          1                   0
## 1242817               0           0            0          1                   1
## 382899                0           0            0          1                   0
## 118951                0           0            1          2                   1
## 855741                0           0            1          1                   1
```

```
## 1551687               0           0           0          2                  1
## 1201431               0           0           0          1                  1
## 471531                0           0           0          4                  1
## 15110                 0           0           0          1                  0
## 1555083               0           0           0          1                  1
## 417764                0           0           0          1                  0
##         loyalty2.loyal prediction correct
## 209931               0       high    TRUE
## 1403545              0        low    TRUE
## 1226812              0        low    TRUE
## 428400               0        low    TRUE
## 417907               1        low    TRUE
## 38464                1        low   FALSE
## 1612610              0       high    TRUE
## 1224527              0        low    TRUE
## 879815               0       high    TRUE
## 318760               1        low    TRUE
## 1242817              0        low    TRUE
## 382899               1       high    TRUE
## 118951               0        low    TRUE
## 855741               0        low    TRUE
## 1551687              0       high    TRUE
## 1201431              0        low    TRUE
## 471531               0        low   FALSE
## 15110                1        low    TRUE
## 1555083              0       high    TRUE
## 417764               1        low    TRUE
```

Overall, our model does a great job at predicting when a purchase will be high profit margin versus low profit margin and gets it right 83% of the time. The aspect of the model that is the least effective is its sensitivity and the aspect that is the most effective is the specificity. This means that the models is very good at classifying the low margin purchases successfully (specificity), but slightly worse at classifying the high margin purchases correctly (sensitivity). Thus, while the model is quite good overall, it particularly excels at identifying bad transactions.