R-Feature Selection

Vivian Bwana

2022-06-10

IDENTIFYING RELEVANT FEATURES USING FEATURE SELECTION TECHNIQUE

Defining the Question

a) Specifying the question

To perform feature selection so as to identify the most relevant features that will influence the marketing strategies and result in the highest no. of sales (total price including tax)

b) Metric for success

To be able to identify the most relevant features that can influence the marketing strategies and result in the highest no. of sales.

c) Understanding the Context

Carrefour is one of the leading retail shops, (supermarkets) in the world. It was founded in France, in 1959. It has over the years expanded it's operations internationaly with the Kenyan branch opening in 1995.It has several branches in different parts of major cities countrywide.

You are a Data analyst at Carrefour Kenya and are currently undertaking a project that will inform the marketing department on the most relevant marketing strategies that will result in the highest no. of sales (total price including tax). Your project has been divided into four parts where you'll explore a recent marketing dataset by performing various unsupervised learning techniques and later providing recommendations based on your insights.

d) Experimental Design

1. Problem Definition
2. Data Sourcing
3. Check the Data
4. Perform Data Cleaning
5. Perform Feature Selection
6. Challenging the Solution
7. Conclusion
8. Recommendation

e) Data Relevance /Sourcing

The dataset is relevant and reliable since it was provided by the client. We were able to draw relevant insights from it.

Data Understanding

Loading Libraries

```
# loading the necessary libraries
library(data.table)
library(caret)
```

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

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

```
library(purrr)
```

```
##
## Attaching package: 'purrr'
```

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

```
## The following object is masked from 'package:data.table':
##
##     transpose
```

```
library(dbplyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:dbplyr':
##
##     ident, sql
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(data.table)
library(latticeExtra)
```

```
##
## Attaching package: 'latticeExtra'
```

```
## The following object is masked from 'package:ggplot2':
##
##     layer
```

Part 2: Feature Selection

This section requires you to perform feature selection through the use of the unsupervised learning methods learned earlier this week. You will be required to perform your analysis and provide insights on the features that contribute the most information to the dataset.

Dataset for Part 2: Feature Selection

```
# loading dataset
library(readr)
cafo12 <- fread("~/Downloads/Supermarket_Dataset_1 - Sales Data.csv")
#preview
head(cafo12)
```

```
##     Invoice ID Branch Customer type Gender           Product line Unit price
##         <char> <char>        <char> <char>                 <char>      <num>
## 1: 750-67-8428      A        Member Female      Health and beauty      74.69
## 2: 226-31-3081      C        Normal Female Electronic accessories      15.28
## 3: 631-41-3108      A        Normal   Male     Home and lifestyle      46.33
## 4: 123-19-1176      A        Member   Male      Health and beauty      58.22
## 5: 373-73-7910      A        Normal   Male      Sports and travel      86.31
## 6: 699-14-3026      C        Normal   Male Electronic accessories      85.39
##    Quantity     Tax      Date   Time     Payment   cogs gross margin percentage
##       <int>   <num>    <char> <char>      <char>  <num>                  <num>
## 1:        7 26.1415  1/5/2019  13:08     Ewallet 522.83               4.761905
## 2:        5  3.8200  3/8/2019  10:29        Cash  76.40               4.761905
## 3:        7 16.2155  3/3/2019  13:23 Credit card 324.31               4.761905
## 4:        8 23.2880 1/27/2019  20:33     Ewallet 465.76               4.761905
## 5:        7 30.2085  2/8/2019  10:37     Ewallet 604.17               4.761905
## 6:        7 29.8865 3/25/2019  18:30     Ewallet 597.73               4.761905
##    gross income Rating    Total
##           <num>  <num>    <num>
## 1:      26.1415    9.1 548.9715
## 2:       3.8200    9.6  80.2200
## 3:      16.2155    7.4 340.5255
## 4:      23.2880    8.4 489.0480
## 5:      30.2085    5.3 634.3785
## 6:      29.8865    4.1 627.6165
```

Exploring the Dataset

Dimensions

```
# checking the dimensions of the datasets
# to see how many rows and coulums there are
dim(cafo12)
```

```
## [1] 1000   16
```

There are 1000 and 16 columns in the first dataset

Data Types

```
#Checking the datatypes of the dataset
str(cafo12)
```

```
## Classes 'data.table' and 'data.frame':   1000 obs. of  16 variables:
##  $ Invoice ID            : chr  "750-67-8428" "226-31-3081" "631-41-3108" "123-19-1176" ...
##  $ Branch                : chr  "A" "C" "A" "A" ...
##  $ Customer type         : chr  "Member" "Normal" "Normal" "Member" ...
##  $ Gender                : chr  "Female" "Female" "Male" "Male" ...
##  $ Product line          : chr  "Health and beauty" "Electronic accessories" "Home and lifestyle" "
##  $ Unit price            : num  74.7 15.3 46.3 58.2 86.3 ...
##  $ Quantity              : int  7 5 7 8 7 7 6 10 2 3 ...
##  $ Tax                   : num  26.14 3.82 16.22 23.29 30.21 ...
##  $ Date                  : chr  "1/5/2019" "3/8/2019" "3/3/2019" "1/27/2019" ...
##  $ Time                  : chr  "13:08" "10:29" "13:23" "20:33" ...
##  $ Payment               : chr  "Ewallet" "Cash" "Credit card" "Ewallet" ...
##  $ cogs                  : num  522.8 76.4 324.3 465.8 604.2 ...
##  $ gross margin percentage: num  4.76 4.76 4.76 4.76 4.76 ...
##  $ gross income          : num  26.14 3.82 16.22 23.29 30.21 ...
##  $ Rating                : num  9.1 9.6 7.4 8.4 5.3 4.1 5.8 8 7.2 5.9 ...
##  $ Total                 : num  549 80.2 340.5 489 634.4 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
#colnames(cafo12)
```

Descriptive Statistics Summary

```
# checking summary of the dataframe
library(Hmisc)
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:caret':
##
##     cluster
```

```
## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
#library(describe)
#describe(cafo12)
summary(cafo12)
```

```
##   Invoice ID           Branch           Customer type         Gender
##  Length:1000        Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Product line         Unit price       Quantity          Tax
##  Length:1000        Min.   :10.08   Min.   : 1.00   Min.   : 0.5085
##  Class :character   1st Qu.:32.88   1st Qu.: 3.00   1st Qu.: 5.9249
##  Mode  :character   Median :55.23   Median : 5.00   Median :12.0880
##                     Mean   :55.67   Mean   : 5.51   Mean   :15.3794
##                     3rd Qu.:77.94   3rd Qu.: 8.00   3rd Qu.:22.4453
##                     Max.   :99.96   Max.   :10.00   Max.   :49.6500
##      Date               Time             Payment              cogs
##  Length:1000        Length:1000        Length:1000        Min.   : 10.17
##  Class :character   Class :character   Class :character   1st Qu.:118.50
##  Mode  :character   Mode  :character   Mode  :character   Median :241.76
##                                                           Mean   :307.59
##                                                           3rd Qu.:448.90
##                                                           Max.   :993.00
##  gross margin percentage  gross income         Rating           Total
##  Min.   :4.762           Min.   : 0.5085   Min.   : 4.000   Min.   :  10.68
##  1st Qu.:4.762           1st Qu.: 5.9249   1st Qu.: 5.500   1st Qu.: 124.42
##  Median :4.762           Median :12.0880   Median : 7.000   Median : 253.85
##  Mean   :4.762           Mean   :15.3794   Mean   : 6.973   Mean   : 322.97
##  3rd Qu.:4.762           3rd Qu.:22.4453   3rd Qu.: 8.500   3rd Qu.: 471.35
##  Max.   :4.762           Max.   :49.6500   Max.   :10.000   Max.   :1042.65
```

The function summary gives the statistical summary of mean, median, minimum, maximum and quantile ranges as shown above

Column Names

```
# checking the column names
#colnames(cafo12)
```

Missing Values

```
#Checking for the sum of Missing values
colSums(is.na(cafo12))
```

```
##               Invoice ID                   Branch            Customer type
##                        0                        0                        0
##                   Gender             Product line               Unit price
##                        0                        0                        0
##                 Quantity                      Tax                     Date
##                        0                        0                        0
##                     Time                  Payment                     cogs
##                        0                        0                        0
## gross margin percentage             gross income                   Rating
##                        0                        0                        0
##                    Total
##                        0
```

Duplicates

```
# checking for duplicates
cafo12.duplicates <- cafo12[duplicated(cafo12),]

#printing duplicated rows
cafo12.duplicates
```

```
## Empty data.table (0 rows and 16 cols): Invoice ID,Branch,Customer type,Gender,Product line,Unit price
```

PERFORMING FEATURE SELECTION

Selecting Numerical Features

Extracting numerical cols to use on Feature selection

```
# extracting numerical columns
nump <-data.frame(cafo12[,c(6,7,8,12,14,15,16)])
# previewing
head(nump)
```

```
##   Unit.price Quantity     Tax   cogs gross.income Rating    Total
## 1      74.69        7 26.1415 522.83      26.1415    9.1 548.9715
## 2      15.28        5  3.8200  76.40       3.8200    9.6  80.2200
## 3      46.33        7 16.2155 324.31      16.2155    7.4 340.5255
## 4      58.22        8 23.2880 465.76      23.2880    8.4 489.0480
## 5      86.31        7 30.2085 604.17      30.2085    5.3 634.3785
## 6      85.39        7 29.8865 597.73      29.8865    4.1 627.6165
```

There are 8 numerical columns

Embedded Methods for Feature Selection

Loading wskm library

```
library(wskm)
```

```
## Loading required package: fpc
```

Preview column names

```
#colnames(nump)
```

Creating the model

```
set.seed(2)
model <- ewkm(nump[,1:7], 3, lambda=2, maxiter=1000)


# extracting numerical columns
#nump <-data.frame(cafo12[,c(6,7,8,12,13,14,15,16)])
# previewing
#head(nump)
```

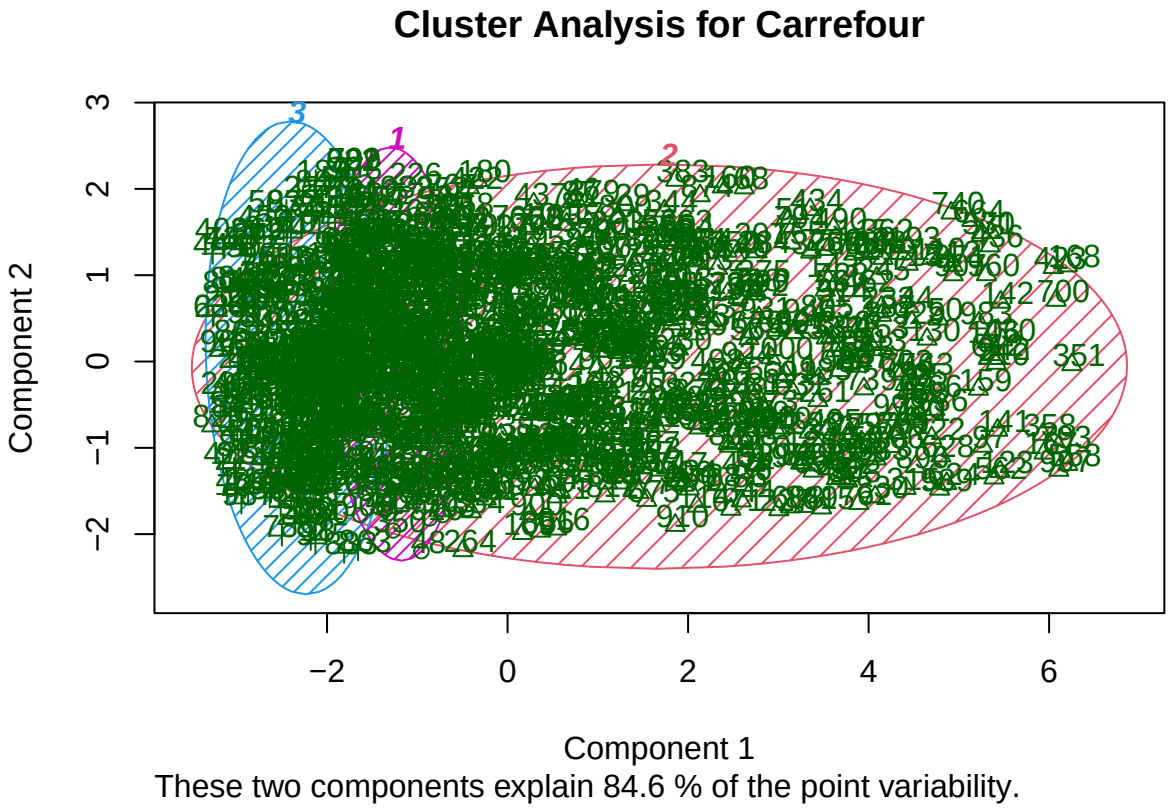
Cluster

```
library("cluster")
```

Plotting the model

```
color= rainbow

clusplot(nump[,1:7], model$cluster, color=TRUE, shade=TRUE,
         labels=2, lines=1,main='Cluster Analysis for Carrefour')
```



Checking stored weights in the model

```
round(model$weights*100,2)
```

```
##   Unit.price Quantity Tax cogs gross.income Rating Total
## 1          0        0  50    0           50   0.00     0
## 2          0        0   0    0            0  99.99     0
## 3          0        0  50    0           50   0.00     0
```

Weights are the measure of the relative importance of each variable with regards to the membership of the observations to that cluster. The weights are incorporated into the distance function, typically reducing the distance for more important variables. Weights remain stored in the model.

Challenging the Solution

Using Filter Method for Feature Selection

Installing and loading Caret package

```
#suppressWarnings(
#        suppressMessages(if
#                         (!require(caret, #quietly=TRUE))
#                 install.packages("caret")))
library(caret)
```

Installing and loading corrplot package

```
#suppressWarnings(
#        suppressMessages(if
#                         (!require(corrplot, #quietly=TRUE))
#                install.packages("corrplot")))
library(corrplot)
```

```
## corrplot 0.92 loaded
```

Correlation matrix

```
# Calculating correlation matrix
correlationMatrix<- cor(nump) #nump[,1:7]

# finding highly correlated attributes
highlyCorrelated <- findCorrelation(correlationMatrix, cutoff=0.75)

# picking highly correlated attributes out
highlyCorrelated
```

```
## [1] 4 7 3
```

```
#Printing names of highly correlated

names(nump[,highlyCorrelated])
```

```
## [1] "cogs"  "Total" "Tax"
```

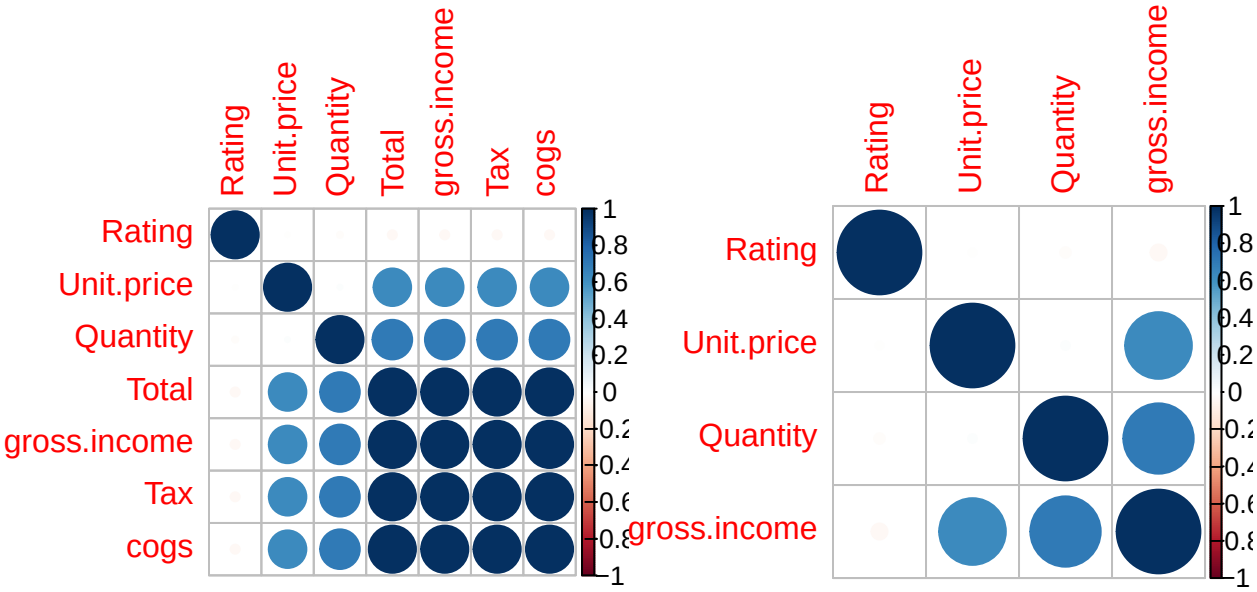
cogs, total and tax have high correlation These will be removed to avoid redundancy

Removing Redundant Features

```
#Removing Redundant Features
nump2<-nump[-highlyCorrelated]
```

Performing our graphical comparison

```
# Plotting
par(mfrow = c(1, 2))
corrplot(correlationMatrix, order = "hclust")
corrplot(cor(nump2), order = "hclust")
```



The correlation matrix looks much better without the highly correlate variables.

Variables that are most important are gross income, quantity and unit price and rating just as was observed in the pca.

Conclusion

Using Embedded Methods for Feature Selection

Generated PC1 and PC2 which indicated that the most important variables are in cluster 2. The weights function shows gives the order of importance as follows: 1. Unit.price 2. Quantity 3. Tax 4. cogs 5. gross.income 6. Rating 7. Total in that order. PCA1 and PCA2 explain 84.6%of the point variability.

Using Filter Method for Feature Selection

The filter method for feature selection shows that cogs, total and tax have high correlation These are removed to avoid redundancy Upon removal, we realize that the variables that are most important are 1. gross income, 2. quantity 3. unit price and 4. rating just as was observed in the pca.

Recommendation

Carrefour should consider these four variables when coming up with a model that is looking into maximizing profit. 1. gross income 2. quantity 3. unit price and 4. rating