GrupoBimboEDA

Anandan Balaji

Contents

1 Grupo Bimbo Exploratory Data Analysis

Objective: Predict the demand based on historical sales data.

1.1 About Dataset & Data fields

The dataset is available at www.kaagle.com

Some of the important data fields:

- Semana — Week number (From Thursday to Wednesday)
- Agencia_ID — Sales Depot ID
- Canal_ID — Sales Channel ID
- Ruta_SAK — Route ID (Several routes = Sales Depot)
- Cliente_ID — Client ID
- NombreCliente — Client name
- Producto_ID — Product ID
- NombreProducto — Product Name
- **Venta_uni_hoy — Sales unit this week** (integer)
- Venta_hoy — Sales this week (unit: pesos)
- Dev_uni_proxima — Returns unit next week (integer)
- Dev_proxima — Returns next week (unit: pesos)
- **Demanda_uni_equil — Adjusted Demand** (integer) (This is the target you will predict)

1.2 Train dataset

```
library("data.table")
system.time(train <- fread("./Dataset/train.csv", header = TRUE))
```

```
##
Read 0.0% of 74180464 rows
Read 3.0% of 74180464 rows
Read 6.0% of 74180464 rows
Read 9.0% of 74180464 rows
Read 11.9% of 74180464 rows
```

```
Read 14.9% of 74180464 rows
Read 17.9% of 74180464 rows
Read 20.9% of 74180464 rows
Read 23.9% of 74180464 rows
Read 26.9% of 74180464 rows
Read 29.9% of 74180464 rows
Read 32.9% of 74180464 rows
Read 35.8% of 74180464 rows
Read 38.8% of 74180464 rows
Read 41.8% of 74180464 rows
Read 44.8% of 74180464 rows
Read 47.8% of 74180464 rows
Read 50.8% of 74180464 rows
Read 53.7% of 74180464 rows
Read 56.7% of 74180464 rows
Read 59.7% of 74180464 rows
Read 62.7% of 74180464 rows
Read 65.7% of 74180464 rows
Read 68.6% of 74180464 rows
Read 71.6% of 74180464 rows
Read 74.6% of 74180464 rows
Read 77.6% of 74180464 rows
Read 80.6% of 74180464 rows
Read 83.5% of 74180464 rows
Read 86.5% of 74180464 rows
Read 89.5% of 74180464 rows
Read 92.5% of 74180464 rows
Read 95.5% of 74180464 rows
Read 98.4% of 74180464 rows
Read 74180464 rows and 11 (of 11) columns from 2.980 GB file in 00:00:39
```

```
##    user  system elapsed
##  37.332   0.860  62.193
```

```
system.time(test <- fread("./Dataset/test.csv", header = TRUE))
```

```
##    user  system elapsed
##   1.628   0.076   2.176
```

```
system.time(product <- fread("./Dataset/producto_tabla.csv", header=TRUE))
```

```
##    user  system elapsed
##   0.000   0.000   0.001
```

```
#structure of train
str(train)
```

```
## Classes 'data.table' and 'data.frame':   74180464 obs. of  11 variables:
##  $ Semana          : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Agencia_ID      : int  1110 1110 1110 1110 1110 1110 1110 1110 1110 1110 ...
##  $ Canal_ID        : int  7 7 7 7 7 7 7 7 7 7 ...
##  $ Ruta_SAK        : int  3301 3301 3301 3301 3301 3301 3301 3301 3301 3301 ...
```

```
##  $ Cliente_ID       : int  15766 15766 15766 15766 15766 15766 15766 15766 15766 15766 ...
##  $ Producto_ID      : int  1212 1216 1238 1240 1242 1250 1309 3894 4085 5310 ...
##  $ Venta_uni_hoy    : int  3 4 4 4 3 5 3 6 4 6 ...
##  $ Venta_hoy        : num  25.1 33.5 39.3 33.5 22.9 ...
##  $ Dev_uni_proxima  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Dev_proxima      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Demanda_uni_equil: int  3 4 4 4 3 5 3 6 4 6 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
## number of observations
nrow(train)
```

```
## [1] 74180464
```

```
## get the weekly data - ie., number of transactions in that particular week.
table(train$Semana)
```

```
##
##        3        4        5        6        7        8        9
## 11165207 11009593 10615397 10191837 10382849 10406868 10408713
```

```
## get the demand info for every week
## tapply(X-vector, Index-variable, function)
##
tapply(train$Demanda_uni_equil, train$Semana, sum)
```

```
##        3        4        5        6        7        8        9
## 77664309 79618866 77610637 73851129 76597014 75525105 75054450
```

```
## number of unque products
length(unique(train$Producto_ID))
```

```
## [1] 1799
```

```
# which is the highest demand product
prod_results <- tapply(train$Demanda_uni_equil, train$Producto_ID, sum)
prod_results <- sort(prod_results, decreasing = TRUE)

highest_demand_prod <- prod_results[1]
highest_demand_prod
```

```
##     2425
## 23728674
```

```
## the most popular product is

str(product)
```

```
## Classes 'data.table' and 'data.frame':   2592 obs. of  2 variables:
##  $ Producto_ID   : int  0 9 41 53 72 73 98 99 100 106 ...
##  $ NombreProducto: chr  "NO IDENTIFICADO 0" "Capuccino Moka 750g NES 9" "Bimbollos Ext sAjonjoli 6p 4
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
product$NombreProducto[2425]
```

```
## [1] "Tortilla Hna RB 10p 260g DH 47840"
```

1.3 Test Dataset

```
# let's look at the test dataset

str(test)
```

```
## Classes 'data.table' and 'data.frame':   6999251 obs. of  7 variables:
##  $ id         : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Semana     : int  11 11 10 11 11 11 11 10 10 11 ...
##  $ Agencia_ID : int  4037 2237 2045 1227 1219 1146 2057 1612 1349 1461 ...
##  $ Canal_ID   : int  1 1 1 1 1 4 1 1 1 1 ...
##  $ Ruta_SAK   : int  2209 1226 2831 4448 1130 6601 4507 2837 1223 1203 ...
##  $ Cliente_ID : int  4639078 4705135 4549769 4717855 966351 1741414 4659766 4414012 397854 1646915 .
##  $ Producto_ID: int  35305 1238 32940 43066 1277 972 1232 35305 1240 43203 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
# look at the week info
table(test$Semana)
```

```
##
##      10      11
## 3538385 3460866
```

```
#look at the products, are all of them available in training dataset?
train_prods <- unique(train$Producto_ID)

test_prods <- unique(test$Producto_ID)

# are the products equal
setequal(train_prods, test_prods)
```

```
## [1] FALSE
```

```
# number of products equal
length(intersect(train_prods, test_prods))
```

```
## [1] 1488
```

```
# get the new products in test dataset
new_prods_in_test <- setdiff(test_prods, train_prods)

# number of new products in test dataset
length(new_prods_in_test)
```

```
## [1] 34
```

```
## look at Agency ID (Sales Depot ID)
setequal(train$Agencia_ID, test$Agencia_ID)
```

```
## [1] TRUE
```

```
## Channel ID (Sales chaneel ID)
setequal(train$Canal_ID, test$Canal_ID)
```

```
## [1] TRUE
```

```
## Route ID ( Ruta_SAK)
setequal(train$Ruta_SAK, test$Ruta_SAK)
```

```
## [1] FALSE
```

```
intersect_routes <- intersect(train$Ruta_SAK, test$Ruta_SAK)

#new routes in test
test_new_routes <- setdiff(test$Ruta_SAK, train$Ruta_SAK)

is.element(test_new_routes[1], test$Ruta_SAK)
```

```
## [1] TRUE
```

```
## examine the client IDs
setequal(train$Cliente_ID, test$Cliente_ID)
```

```
## [1] FALSE
```

```
new_test_clients <- setdiff(test$Cliente_ID, train$Cliente_ID)

# how many new clients?
length(new_test_clients)
```

```
## [1] 9663
```

```
is.element(new_test_clients[1], test$Cliente_ID)
```

```
## [1] TRUE
```