bimbo_features.R

gquai

2020-01-15

```
# DATA SCIENCE ACADEMY
# Big Data Analytics com R e Microsoft Azure Machine Learning
#
# Model to accurately predict inventory demand based on data sales histories
#
# Gabriel Quaiotti
# Dez 2019
#
# In this competition, you will forecast the demand of a product for a given week, at a
# particular store.
# The dataset you are given consists of 9 weeks of sales transactions in Mexico. Every week,
# there are delivery
# trucks that deliver products to the vendors. Each transaction consists of sales and returns.
# Returns are the products that are unsold and expired. The demand for a product in a certain
# week is defined as the sales this week subtracted by the return next week.
#
#
# The train and test dataset are split based on time, as well as the public and private
# leaderboard dataset split.
#
#
# Things to note:
#
# There may be products in the test set that don't exist in the train set. This is the expected
# behavior of inventory data,
# since there are new products being sold all the time. Your model should be able to accommodate
# this.
#
# The adjusted demand (Demanda_uni_equil) is always >= 0 since demand should be either 0 or a
# positive value. The reason that Venta_uni_hoy - Dev_uni_proxima
# sometimes has negative values is that the returns records sometimes carry over a few weeks.

# File descriptions
# train.csv - the training set
# test.csv - the test set
# sample_submission.csv - a sample submission file in the correct format
# cliente_tabla.csv - client names (can be joined with train/test on Cliente_ID)
# producto_tabla.csv - product names (can be joined with train/test on Producto_ID)
# town_state.csv - town and state (can be joined with train/test on Agencia_ID)

# 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)

setwd('D:/Github/DSA_BIMBO_INVENTORY')

library(data.table)
library(caTools)
library(caret)
```

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

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

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(lares)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## Registered S3 methods overwritten by 'forecast':
##   method             from
##   fitted.fracdiff    fracdiff
##   residuals.fracdiff fracdiff
```

```
library(MASS)

v_c_file_train <- "dataset/train_clean.csv"
v_c_file_out <- "dataset/train_features.csv"

gc()
```

```
##           used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells 2716138 145.1    4102610 219.2  4102610 219.2
## Vcells 4310940  32.9   10146329  77.5  7140074  54.5
```

```
###################################
# TRAIN DATASET
###################################
train <- data.table::fread(file = v_c_file_train)

# Balance by product and customer
df <- train[, list(prod_cust_balance = median(Demanda_uni_equil), prod_cust_bal_mean = mean(Demanda_uni_
train <- merge(x = train, y = df, all.x = TRUE, by.x = c("Cliente_ID", "Producto_ID"), by.y = c("Client

# Balance by route and product
df <- train[, list(prod_rout_balance = median(Demanda_uni_equil)), by = list(Ruta_SAK, Producto_ID)]
train <- merge(x = train, y = df, all.x = TRUE, by.x = c("Ruta_SAK", "Producto_ID"), by.y = c("Ruta_SAK

# Balance by route and customer
df <- train[, list(cust_rout_balance = median(Demanda_uni_equil)), by = list(Ruta_SAK, Cliente_ID)]
train <- merge(x = train, y = df, all.x = TRUE, by.x = c("Ruta_SAK", "Cliente_ID"), by.y = c("Ruta_SAK"

# Route Size
df1 <- train[ , list(rout_sum = sum(Demanda_uni_equil)), by = list(Ruta_SAK)]

# Product proportion at the route (when we not know the customer this may be useful)
df <- train[ , list(prod_sum = sum(Demanda_uni_equil)), by = list(Ruta_SAK, Producto_ID)]
df <- merge(x = df, y = df1, all.x = TRUE, by.x = c("Ruta_SAK"), by.y = c("Ruta_SAK"))
df[rout_sum > 0, prod_rout_prop := prod_sum / rout_sum]
train <- merge(x = train, y = df[,list(Ruta_SAK, Producto_ID, prod_rout_prop)], all.x = TRUE, by.x = c(

# Customer proportion at the route (when we not know the product this may be useful)
df <- train[ , list(cust_sum = sum(Demanda_uni_equil)), by = list(Ruta_SAK, Cliente_ID)]
df <- merge(x = df, y = df1, all.x = TRUE, by.x = c("Ruta_SAK"), by.y = c("Ruta_SAK"))
df[rout_sum > 0, cust_rout_prop := cust_sum / rout_sum]
train <- merge(x = train, y = df[,list(Ruta_SAK, Cliente_ID, cust_rout_prop)], all.x = TRUE, by.x = c("

remove(df1)
remove(df)

# Order by Cliente_ID, Producto_ID, Semana
train <- data.table::setorder(x = train, Cliente_ID, Producto_ID, Semana)

# Get the balance from previous line
train[,last_week_balance := shift(x = Demanda_uni_equil, n = 1)]
train[!(Cliente_ID == shift(Cliente_ID, n = 1) & Producto_ID == shift(Producto_ID, n = 1)), last_week_ba
train[is.na(last_week_balance), last_week_balance := 0]

train[, Agencia_ID := NULL]
train[, Canal_ID := NULL]
train[, Venta_uni_hoy := NULL]
train[, Venta_hoy := NULL]
train[, Dev_uni_proxima := NULL]
train[, Dev_proxima := NULL]
train[, previous_week_balance := NULL]
```

```
## Warning in `[.data.table`(train, , `:=`(previous_week_balance, NULL)): Column
## 'previous_week_balance' does not exist to remove
```

```r
train[, unit_price := NULL]
train[, dev_unit_price := NULL]

fwrite(x = train, file = v_c_file_out)

rm(list=ls())
gc()
```

```
##            used  (Mb) gc trigger    (Mb)   max used    (Mb)
## Ncells  2753804 147.1    4102610   219.2    4102610   219.2
## Vcells 67912979 518.2 1711818331 13060.2 2139665969 16324.4
```