Prevendo Demanda de Estoque com Base em Vendas

Luiz Carlos Castro Cunha Junior

2019-11-10

```
# Definindo  o working directory
setwd("C:/FCD/BigDataAnalytics-R-Azure/Projeto2")

# Carregando o pacote para manipulação de dados
library(tidyverse)
```

```
## -- Attaching packages ------------------------ tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts --------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(rmarkdown)

# Carregando os dados
dataset<- read_csv("dataset/train.csv", n_max = 15000000)
```

```
## Parsed with column specification:
## cols(
##   Semana = col_double(),
##   Agencia_ID = col_double(),
##   Canal_ID = col_double(),
##   Ruta_SAK = col_double(),
##   Cliente_ID = col_double(),
##   Producto_ID = col_double(),
##   Venta_uni_hoy = col_double(),
##   Venta_hoy = col_double(),
##   Dev_uni_proxima = col_double(),
##   Dev_proxima = col_double(),
##   Demanda_uni_equil = col_double()
## )
```

```
#Visualizando os dados
head(dataset)
```

```
## # A tibble: 6 x 11
##   Semana Agencia_ID Canal_ID Ruta_SAK Cliente_ID Producto_ID Venta_uni_hoy
##    <dbl>      <dbl>    <dbl>    <dbl>      <dbl>       <dbl>         <dbl>
## 1      3       1110        7     3301      15766        1212             3
## 2      3       1110        7     3301      15766        1216             4
## 3      3       1110        7     3301      15766        1238             4
## 4      3       1110        7     3301      15766        1240             4
## 5      3       1110        7     3301      15766        1242             3
## 6      3       1110        7     3301      15766        1250             5
## # ... with 4 more variables: Venta_hoy <dbl>, Dev_uni_proxima <dbl>,
```

```
## #   Dev_proxima <dbl>, Demanda_uni_equil <dbl>
```

```
str(dataset)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 15000000 obs. of  11 variables:
##  $ Semana           : num  3 3 3 3 3 3 3 3 3 3 ...
##  $ Agencia_ID       : num  1110 1110 1110 1110 1110 1110 1110 1110 1110 1110 ...
##  $ Canal_ID         : num  7 7 7 7 7 7 7 7 7 7 ...
##  $ Ruta_SAK         : num  3301 3301 3301 3301 3301 ...
##  $ Cliente_ID       : num  15766 15766 15766 15766 15766 ...
##  $ Producto_ID      : num  1212 1216 1238 1240 1242 ...
##  $ Venta_uni_hoy    : num  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  : num  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: num  3 4 4 4 3 5 3 6 4 6 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Semana = col_double(),
##   ..   Agencia_ID = col_double(),
##   ..   Canal_ID = col_double(),
##   ..   Ruta_SAK = col_double(),
##   ..   Cliente_ID = col_double(),
##   ..   Producto_ID = col_double(),
##   ..   Venta_uni_hoy = col_double(),
##   ..   Venta_hoy = col_double(),
##   ..   Dev_uni_proxima = col_double(),
##   ..   Dev_proxima = col_double(),
##   ..   Demanda_uni_equil = col_double()
##   .. )
```

```
summary(dataset)
```

```
##      Semana        Agencia_ID       Canal_ID         Ruta_SAK
##  Min.   :3.000   Min.   : 1110   Min.   : 1.000   Min.   :   1
##  1st Qu.:3.000   1st Qu.: 1220   1st Qu.: 1.000   1st Qu.:1175
##  Median :3.000   Median : 1349   Median : 1.000   Median :1411
##  Mean   :3.256   Mean   : 2149   Mean   : 1.338   Mean   :2069
##  3rd Qu.:4.000   3rd Qu.: 1968   3rd Qu.: 1.000   3rd Qu.:2162
##  Max.   :4.000   Max.   :25759   Max.   :11.000   Max.   :9975
##    Cliente_ID         Producto_ID    Venta_uni_hoy        Venta_hoy
##  Min.   :2.600e+01   Min.   :   41   Min.   :   0.000   Min.   :     0.0
##  1st Qu.:3.503e+05   1st Qu.: 1240   1st Qu.:   2.000   1st Qu.:    16.8
##  Median :1.140e+06   Median : 9217   Median :   3.000   Median :    30.0
##  Mean   :1.746e+06   Mean   :19755   Mean   :   7.048   Mean   :    66.7
##  3rd Qu.:2.323e+06   3rd Qu.:36748   3rd Qu.:   6.000   3rd Qu.:    56.2
##  Max.   :2.015e+09   Max.   :49997   Max.   :5000.000   Max.   :382694.4
##  Dev_uni_proxima      Dev_proxima        Demanda_uni_equil
##  Min.   :    0.000   Min.   :     0.00   Min.   :   0.000
##  1st Qu.:    0.000   1st Qu.:     0.00   1st Qu.:   2.000
##  Median :    0.000   Median :     0.00   Median :   3.000
##  Mean   :    0.113   Mean   :     1.12   Mean   :   6.972
##  3rd Qu.:    0.000   3rd Qu.:     0.00   3rd Qu.:   6.000
##  Max.   :16345.000   Max.   :130760.00   Max.   :5000.000
```

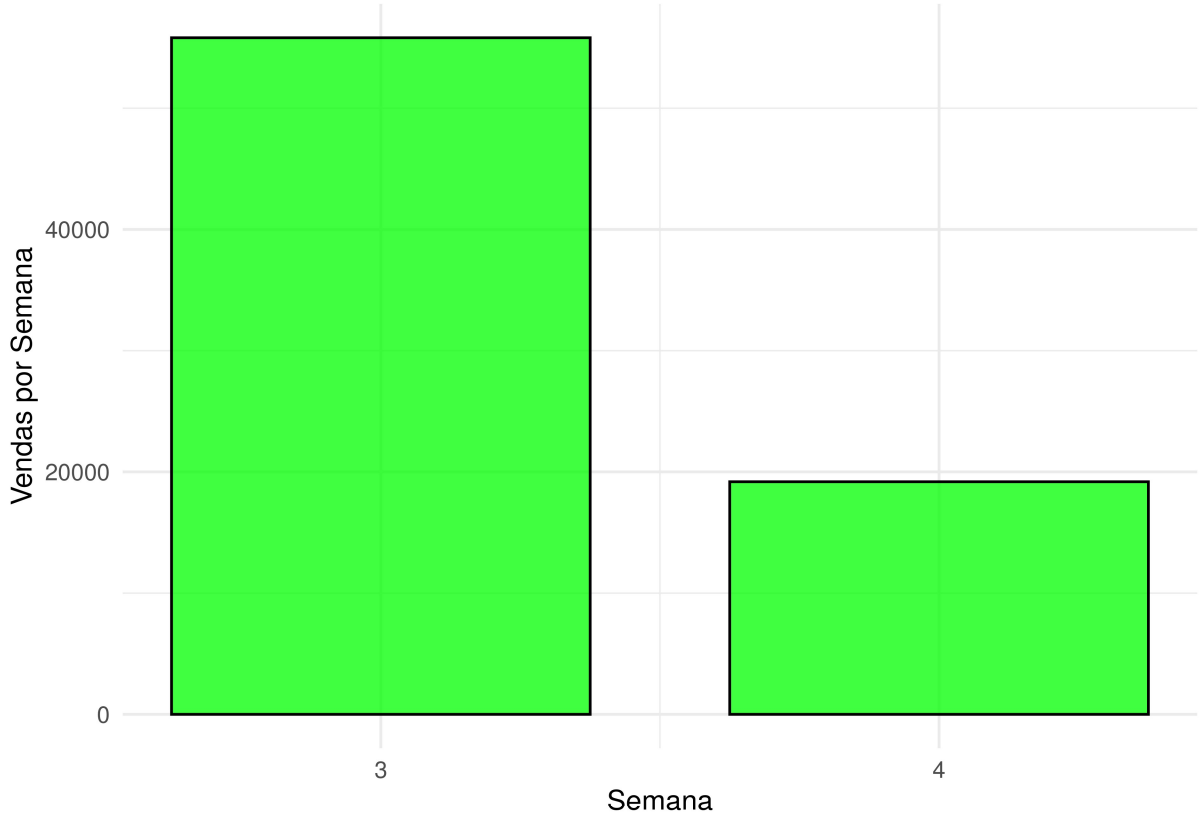
```
#Verificando se há valores NA
colSums(is.na(dataset))
```

```
##             Semana        Agencia_ID          Canal_ID          Ruta_SAK
##                  0                 0                 0                 0
##         Cliente_ID       Producto_ID     Venta_uni_hoy         Venta_hoy
##                  0                 0                 0                 0
##    Dev_uni_proxima       Dev_proxima Demanda_uni_equil
##                  0                 0                 0
```

```
#Verificando os dias de semana
table(dataset$Semana)
```

```
##
##        3        4
## 11165207  3834793
```

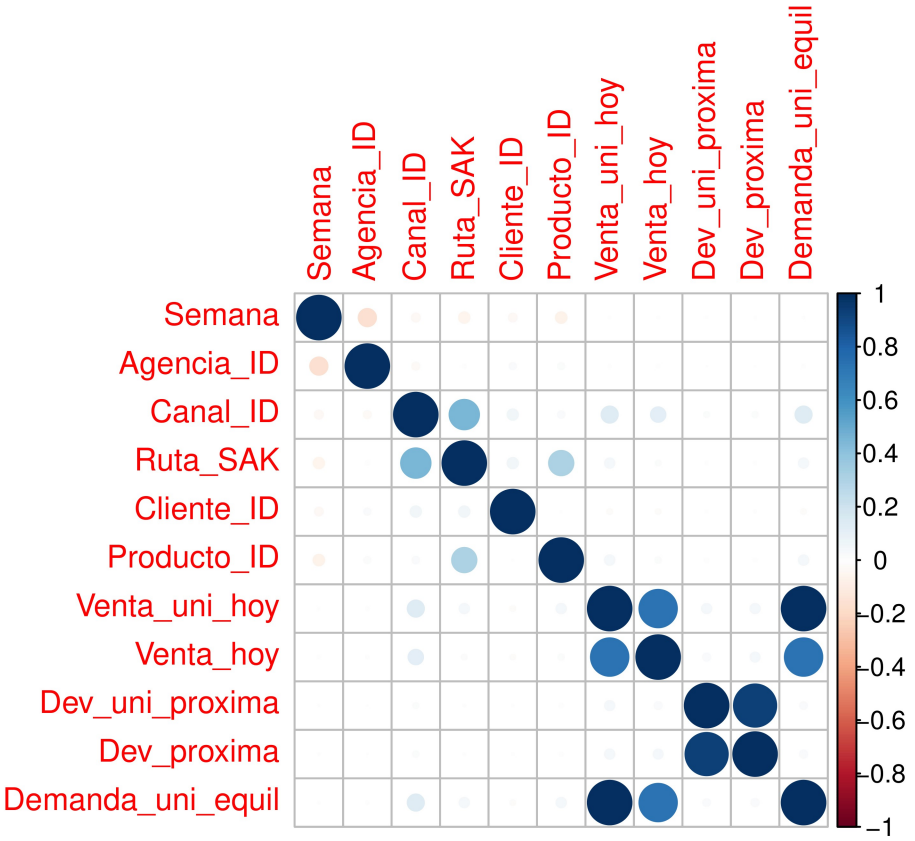
```
#Visualizando a relação de vendas por semana
ggplot(data = dataset %>% sample_frac(0.005)) +
  geom_bar(mapping = aes(x =Semana), alpha = 0.75, color = "black", fill = "green", width = 0.75)+
  scale_x_continuous(breaks = 2:5) +
  scale_y_continuous("Vendas por Semana") +
  theme_minimal()
```




```
#Visualizando a correlação entre as variáveis
library(corrplot)
```

```
## corrplot 0.84 loaded
```

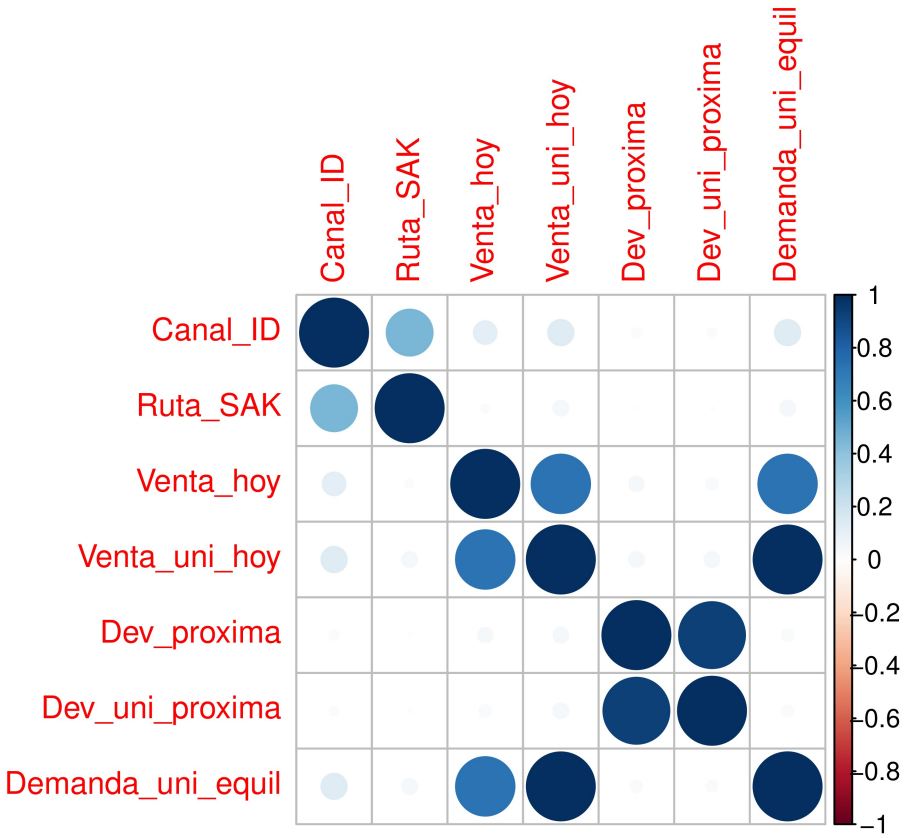
```
corrplot(cor(dataset))
```




```
#Criando um subset com as variáveis que possuem correlação com a variável target
dataset_treino <- dataset %>%
  select(Canal_ID, Ruta_SAK, Venta_hoy, Venta_uni_hoy, Dev_proxima,Dev_uni_proxima, Demanda_uni_equil)

rm(dataset)

#Verificando a correlação entre as variáveis do subset
corrplot(cor(dataset_treino))
```




```
#Separando os dados em treino e teste
library(caret)
```

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

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

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

```
set.seed(5000)

split <- createDataPartition(y = dataset_treino$Canal_ID, p = 0.70, list = F)

treino <- dataset_treino[split,]
teste <- dataset_treino[-split,]

#Verificando a proporção dos dados
nrow(teste) + nrow(treino) == nrow(dataset_treino)
```

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

```
rm(dataset_treino)
rm(split)

#Criando o modelo
set.seed(1234)
```

```r
modelo <- lm(Demanda_uni_equil ~ ., data=treino)

#Verificando o desempenho do modelo
summary(modelo)
```

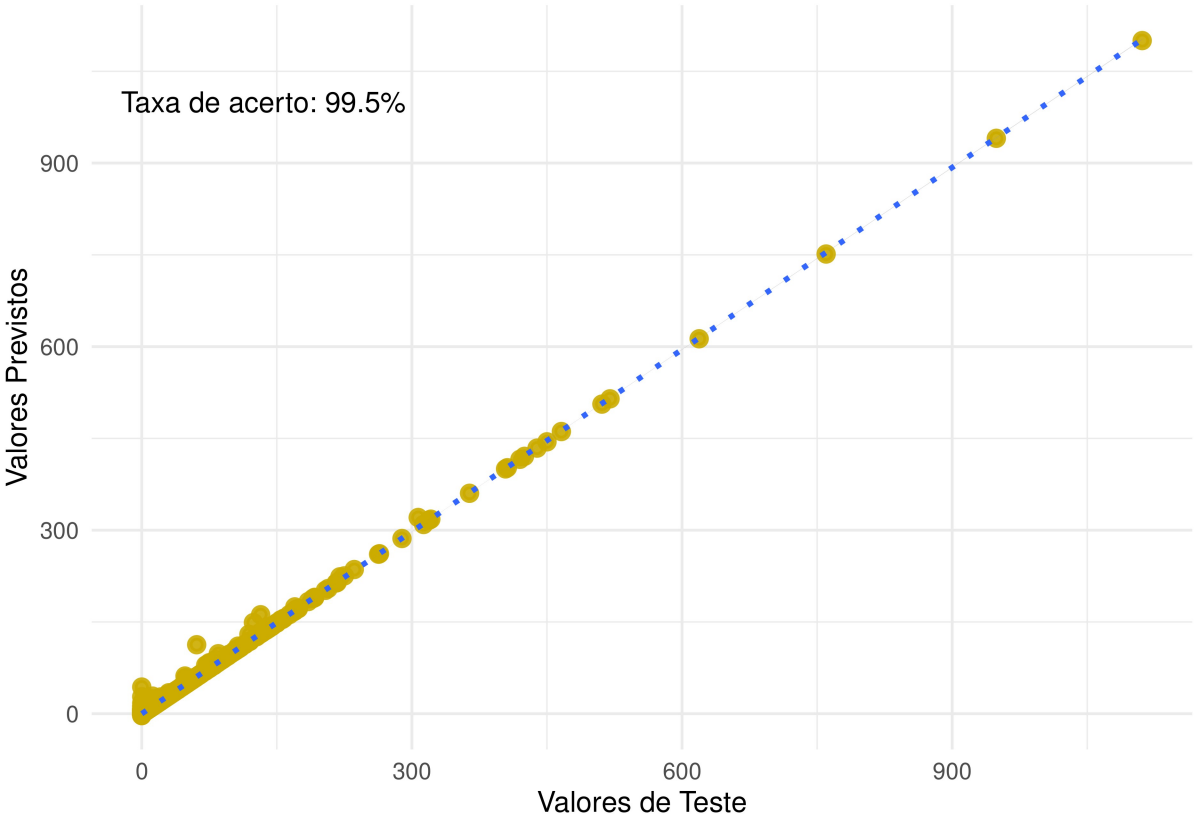
```
##
## Call:
## lm(formula = Demanda_uni_equil ~ ., data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2022.91     0.01     0.03     0.07  1194.50
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     -1.402e-02  7.993e-04   -17.55   <2e-16 ***
## Canal_ID        -8.194e-03  3.676e-04   -22.29   <2e-16 ***
## Ruta_SAK         1.582e-05  3.482e-07    45.42   <2e-16 ***
## Venta_hoy        2.099e-04  2.035e-06   103.13   <2e-16 ***
## Venta_uni_hoy    9.868e-01  3.058e-05 32273.00   <2e-16 ***
## Dev_proxima      1.788e-02  2.927e-05   610.66   <2e-16 ***
## Dev_uni_proxima -2.161e-01  2.464e-04  -877.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.429 on 10499993 degrees of freedom
## Multiple R-squared:  0.9954, Adjusted R-squared:  0.9954
## F-statistic: 3.816e+08 on 6 and 10499993 DF,  p-value: < 2.2e-16
```

```r
previsoes <- predict(modelo, teste)

score<- data.frame(valor_teste = teste$Demanda_uni_equil,
                   valor_previsto = previsoes)


#Relação entre os valores previstos e o valor de teste

ggplot(data = score %>% sample_frac(0.005), aes(x = valor_teste, y= valor_previsto))+
  geom_point(stroke = 1.5, color = 'gold3', alpha = 0.9)+
  geom_smooth(method = 'lm', linetype = 3)+
  annotate("text", x=135, y=1000, label= "Taxa de acerto: 99.5%") +
  labs(x = 'Valores de Teste', y = 'Valores Previstos', main = 'Valores de Teste x Valores Previstos')+
  theme_minimal()
```


Taxa de acerto: 99.5%
Valores Previstos
0
300
600
900
Valores de Teste
0
300
600
900