Sales Performance

Eka Pramudita

5/24/2021

Choose Library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

Get Data

```
data <- read.csv("Raw Data - 4.csv")
sapply(data, class) # check data types
```

```
##                             ID                      join.date
##                      "integer"                    "character"
##                  date.of.birth                         gender
##                    "character"                    "character"
##                           city                GMV..in.scale.
##                    "character"                    "character"
##                Count.of.Invoice             Count.of.Customer
```

```
##                         "integer"                         "integer"
##               X.GMV.of.cigarettes Total.Voucher.Discount..in.scale.
##                         "character"                       "character"
##              Count.of.active.days          Count.of.cancelled.order
##                         "integer"                         "integer"
##                 Profit..in.scale.             Commission..in.scale.
##                         "character"                       "character"
##                       Count.of.SKU                 Count.of.Category
##                         "integer"                         "integer"
```

Data Cleaning

Change numerical columns that are still being character.

```
data$GMV..in.scale. <-
  as.double(gsub(',', '', data$GMV..in.scale.))
data$X.GMV.of.cigarettes <-
  as.double(gsub('%', '', data$X.GMV.of.cigarettes))/100
data$Total.Voucher.Discount..in.scale. <-
  as.double(gsub(',', '', data$Total.Voucher.Discount..in.scale.))
data$Profit..in.scale. <-
  as.double(gsub(',', '', data$Profit..in.scale.))
data$Commission..in.scale. <-
  as.double(gsub(',', '', data$Commission..in.scale.))
```

Data Slicing and Scaling

Select variables that will be included in the initialization model. The selected variables then will be scaled so that we can measure the effect of each variables.

```
scaled.data <- scale(data[c(7:15)], center = TRUE, scale = TRUE)
scaled.data <- cbind(data[6], as.data.frame(scaled.data))
```

Multiple Linear Regression (Initialization)

The scaled data will be modeled using multiple linear regression method.

```
model <- lm(GMV..in.scale. ~ Count.of.Invoice + Count.of.Customer
            + X.GMV.of.cigarettes + Total.Voucher.Discount..in.scale.
            + Count.of.active.days + Count.of.cancelled.order
            + Profit..in.scale. + Commission..in.scale. + Count.of.SKU,
            data = scaled.data)
summary(model)
```

```
##
## Call:
## lm(formula = GMV..in.scale. ~ Count.of.Invoice + Count.of.Customer +
##     X.GMV.of.cigarettes + Total.Voucher.Discount..in.scale. +
##     Count.of.active.days + Count.of.cancelled.order + Profit..in.scale. +
##     Commission..in.scale. + Count.of.SKU, data = scaled.data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -285902  -30021    2477   23745  259804
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         92653       8135  11.390  < 2e-16 ***
## Count.of.Invoice                   -58260      33769  -1.725  0.08981 .
## Count.of.Customer                  -89707      44918  -1.997  0.05051 .
## X.GMV.of.cigarettes                 21753       9479   2.295  0.02538 *
## Total.Voucher.Discount..in.scale.  -21975      24238  -0.907  0.36833
## Count.of.active.days                59799      23141   2.584  0.01230 *
## Count.of.cancelled.order             1895      11120   0.170  0.86525
## Profit..in.scale.                  136852      50460   2.712  0.00878 **
## Commission..in.scale.              163849      51970   3.153  0.00256 **
## Count.of.SKU                       -78808      27739  -2.841  0.00619 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67080 on 58 degrees of freedom
## Multiple R-squared:  0.8722, Adjusted R-squared:  0.8524
## F-statistic: 43.99 on 9 and 58 DF,  p-value: < 2.2e-16
```

We can see from the Pr(>|t|) or p-value that there's still some variables that are not significant to the model. To find the optimal model, Stepwise Regression will be used.

Stepwise Regression

```
ols_step_both_p(model, pent = 0.05, prem = 0.3, details = FALSE)
```

```
##
##                                        Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------------
##                                     Added/                   Adj.
## Step          Variable             Removed     R-Square    R-Square      C(p)         AIC          RMSE
## ----------------------------------------------------------------------------------------------------
##    1    Commission..in.scale.     addition       0.627       0.622    105.1480    1772.3984    107388
##    2      Count.of.Invoice        addition       0.781       0.775     37.2300    1738.1315     82882
##    3     X.GMV.of.cigarettes      addition       0.829       0.821     17.8060    1723.5928     73963
##    4      Profit..in.scale.       addition       0.842       0.832     13.5890    1719.9297     71507
##    5      Count.of.Customer       addition       0.852       0.841     10.9640    1717.3887     69714
## ----------------------------------------------------------------------------------------------------
```

```
model1 <- lm(GMV..in.scale. ~ Count.of.Invoice + Count.of.Customer
             + X.GMV.of.cigarettes + Profit..in.scale.
             + Commission..in.scale.,
            data = scaled.data)
summary(model1)
```

```
##
```

```
## Call:
## lm(formula = GMV..in.scale. ~ Count.of.Invoice + Count.of.Customer +
##     X.GMV.of.cigarettes + Profit..in.scale. + Commission..in.scale.,
##     data = scaled.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -295250  -28959    2747   23791  305435
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              92653       8454  10.960 3.84e-16 ***
## Count.of.Invoice        -66529      22176  -3.000 0.003884 **
## Count.of.Customer       -69496      33586  -2.069 0.042700 *
## X.GMV.of.cigarettes      33819       8909   3.796 0.000337 ***
## Profit..in.scale.       106495      46996   2.266 0.026950 *
## Commission..in.scale.   138154      47458   2.911 0.004999 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69710 on 62 degrees of freedom
## Multiple R-squared:  0.8525, Adjusted R-squared:  0.8406
## F-statistic: 71.65 on 5 and 62 DF,  p-value: < 2.2e-16
```

As we can see now we have model that all variables are significant with 0.05 significance level. If we want to measure importance, we can see the absolute value of coefficient. From the summary we can see that Commission become the most important variable since the absolute coefficient is the largest among all. Then the least important is % GMV of Cigarettes as it has the smallest absolute coefficient among all.