Attribution Analysis of Nike Products' Sales

By Xiaoxuan Ma

In this project, I used stepwise OLS regression to identify how the product attributes affect product sales. And I interpreted the results in combination with commercial scenarios. The raw data is provided by Nike in 2020 and is modified for interview tests. And the raw data won't be shared to others.

Contents

Business Question Overview

The **'RAWDATA'** contains the data of sold quantity of products, and products' attributes, like 'Number of products with the same style', 'If the color is Black Gold'. The attributes include interval variables, dummy variables, categorical variable.

We need to identify how the product attributes affect product sales. Basing on attributes to predict product sales is a common use of machine learning, like random forest. But we need to identify the specific influence. It sounds like we should consider the regression model first, because the coefficients of regression model clearly show the influence of attributes.

Data Cleaning & Preprocessing

1. Missing Value

First, I check if there is any NaN in the data.

```
SKU                          0
SALES_QTY                    0
IF_CORE_PRICE                0
IF_PREMIUM_PRICE             0
IF_GENERAL_PRICE             0
NUM_SAME_PRICE_IN1PE         0
NUM_SAME_PRICE_IN1PE1CAT     0
SSNRK_1ST_SKU_LAUNCH         0
SEASON_SINCE_1ST_SKU_LAUNCH  0
NUM_COLOR_DESC               0
COLOR_CD_1                   0
COLOR_CD_2                   0
COLOR_CD_3                   0
IF_FW_BLACKGOLD              0
IF_TRIPLEWHITE               0
IF_GREY                      0
IF_UNIVRED                   0
IF_BLACK_WHITE               0
IF_BRAND_STORY               0
IF_RETRO                     0
IF_SEASONAL_SILO_SU          0
IF_SEASONAL_SILO_HO          0
NUMSKU_SAME_STYLE_IN1SSN     0
NUMSKU_SAME_MODEL_IN1SSN     0
NUMSKU_SAME_PF_IN1SSN        0
LAST_SEASON_INDICATOR        0
IF_NEW_NEW                   0
IF_NEW_SEASONAL              0
IF_CARRYOVER                 0
OMD_DAYS_SINCE_SEASON_BEGIN  0
```

There isn't.

2. Constant Variable

I found numbers in column 'NUM_COLOR_DESC' are 2, calculate the column's standard deviation, if it equals 0, drop that column. I found 5 constant columns: 'NUM_COLOR_DESC','IF_GREY','IF_UNIVRED','IF_BRAND_STORY','IF_SEASONAL_SILO_HO'

3. Categorical variables convert to dummy variables

Columns 'COLOR_CD_1', 'COLOR_CD_2' and 'COLOR_CD_3' are Categorical variables. I used pd.get_dummies() function to convert them to dummy variables.

4. Data type transfer

When reading the Excel file, the dummy variables' data type is 'int64', use astype() function to change it into 'int8'.

Then, I got a 439 * 50 dataframe with 49 variables

5. Correlation coefficient overview

Index	SALES_QTY
SALES_QTY	1
IF_CORE_PRICE	-0.0613449
IF_PREMIUM_PRICE	0.0453736
IF_GENERAL_PRICE	0.0815874
NUM_SAME_PRICE_IN1PE	-0.0493884
NUM_SAME_PRICE_IN1PE1CAT	0.0290441
SSNRK_1ST_SKU_LAUNCH	-0.226849
SEASON_SINCE_1ST_SKU_LAUNCH	0.226849
IF_FW_BLACKGOLD	-0.0132544
IF_TRIPLEWHITE	0.175803
IF_BLACK_WHITE	-0.015959
IF_RETRO	0.223424
IF_SEASONAL_SILO_SU	-0.0795709
NUMSKU_SAME_STYLE_IN1SSN	0.0851123
NUMSKU_SAME_MODEL_IN1SSN	0.24898
NUMSKU_SAME_PF_IN1SSN	0.26678
LAST_SEASON_INDICATOR	0.0697803
IF_NEW_NEW	-0.10572
IF_NEW_SEASONAL	0.0809177
IF_CARRYOVER	0.0697803
OMD_DAYS_SINCE_SEASON_BEGIN	-0.030705
color1_0	0.000528924
color1_1	0.180419
color1_2	0.0313387
color1_3	-0.0697835
color1_4	-0.0820256

Index	SALES_QTY
color1_5	-0.0299152
color1_6	-0.0435041
color1_7	-0.0470972
color1_8	-0.0593837
color1_9	-0.00366117
color2_0	-0.104003
color2_1	0.0584653
color2_2	-0.0314637
color2_3	0.00511635
color2_4	0.0500417
color2_5	0.104415
color2_6	0.0874451
color2_7	-0.0262735
color2_8	-0.0215852
color3_0	0.0271802
color3_1	-0.0201067
color3_2	0.0743627
color3_3	-0.0646351
color3_4	0.0443985
color3_5	-0.0403274
color3_6	0.00626529
color3_7	-0.00979031
color3_8	-0.0459101
color3_9	-0.0491349

This list shows the correlation coefficient of SALES_QTY and each variable, most of them have a very small absolute value. That is to say, most of them have little impact on SALES_QTY. To more robustly and specifically identify their impact, I used OLS regression.

OLS regression

1. Train test split

Divide data in to two parts, one for training, one for testing, ratio is 8:2, set random_state =11

2. All attributes regression

Now, use all attributes as independent variables to fit the OLS Regression model with SALES_QTY as dependent variables. And I tried two ways of regression: with intercept, without intercept.
Model results:

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              SALES_QTY   R-squared:                       0.282
Model:                            OLS   Adj. R-squared:                  0.184
Method:                 Least Squares   F-statistic:                     2.882
Date:                Sun, 12 Jan 2020   Prob (F-statistic):           8.69e-08
Time:                        18:24:39   Log-Likelihood:                -2237.4
No. Observations:                 351   AIC:                             4561.
Df Residuals:                     308   BIC:                             4727.
Df Model:                          42                                         
Covariance Type:            nonrobust                                         
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
IF_CORE_PRICE                  -2.7743     25.920     -0.107      0.915     -53.777      48.229
IF_PREMIUM_PRICE               20.0452     27.635      0.725      0.469     -34.332      74.422
IF_GENERAL_PRICE               40.0465     20.030      1.999      0.046       0.633      79.460
NUM_SAME_PRICE_IN1PE           -0.2569      0.170     -1.516      0.131      -0.590       0.077
NUM_SAME_PRICE_IN1PE1CAT        0.1370      0.194      0.705      0.481      -0.245       0.519
SSNRK_1ST_SKU_LAUNCH            2.3586      2.399      0.983      0.326      -2.362       7.079
SEASON_SINCE_1ST_SKU_LAUNCH    10.3559      4.919      2.105      0.036       0.677      20.035
IF_FW_BLACKGOLD              1.089e-11   1.23e-11      0.887      0.376   -1.33e-11    3.51e-11
IF_TRIPLEWHITE                213.7670    140.347      1.523      0.129     -62.393     489.927
IF_BLACK_WHITE                  2.8193     60.804      0.046      0.963    -116.825     122.464
IF_RETRO                      357.6266     58.932      6.068      0.000     241.666     473.588
IF_SEASONAL_SILO_SU          -116.0929     44.658     -2.600      0.010    -203.967     -28.219
NUMSKU_SAME_STYLE_IN1SSN       16.0496      7.027      2.284      0.023       2.222      29.878
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              SALES_QTY   R-squared:                       0.282
Model:                            OLS   Adj. R-squared:                  0.184
Method:                 Least Squares   F-statistic:                     2.882
Date:                Sun, 12 Jan 2020   Prob (F-statistic):           8.69e-08
Time:                        18:25:44   Log-Likelihood:                -2237.4
No. Observations:                 351   AIC:                             4561.
Df Residuals:                     308   BIC:                             4727.
Df Model:                          42                                         
Covariance Type:            nonrobust                                         
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                           0.4876      0.229      2.133      0.034       0.038       0.937
IF_CORE_PRICE                  -2.7743     25.920     -0.107      0.915     -53.777      48.229
IF_PREMIUM_PRICE               20.0452     27.635      0.725      0.469     -34.332      74.422
IF_GENERAL_PRICE               40.0465     20.030      1.999      0.046       0.633      79.460
NUM_SAME_PRICE_IN1PE           -0.2569      0.170     -1.516      0.131      -0.590       0.077
NUM_SAME_PRICE_IN1PE1CAT        0.1370      0.194      0.705      0.481      -0.245       0.519
SSNRK_1ST_SKU_LAUNCH            2.3398      2.394      0.977      0.329      -2.371       7.050
SEASON_SINCE_1ST_SKU_LAUNCH    10.3372      4.911      2.105      0.036       0.674      20.000
IF_FW_BLACKGOLD              1.328e-15   8.22e-14      0.016      0.987    -1.6e-13    1.63e-13
IF_TRIPLEWHITE                213.7670    140.347      1.523      0.129     -62.393     489.927
IF_BLACK_WHITE                  2.8193     60.804      0.046      0.963    -116.825     122.464
IF_RETRO                      357.6266     58.932      6.068      0.000     241.666     473.588
IF_SEASONAL_SILO_SU          -116.0929     44.658     -2.600      0.010    -203.967     -28.219
NUMSKU_SAME_STYLE_IN1SSN       16.0496      7.027      2.284      0.023       2.222      29.878
```

Both model have low R-squared value, generally, these model can not interpret sample data very well. But we can find some attributes are statistically significant, which means they can interpret part of the sample data. According to two models' results, I found six significant attributes:

'IF_GENERAL_PRICE',
'SEASON_SINCE_1ST_SKU_LAUNCH',
'IF_RETRO',
'IF_SEASONAL_SILO_SU',
'NUMSKU_SAME_STYLE_IN1SSN',
'OMD_DAYS_SINCE_SEASON_BEGIN'

3. Part of attributes regression

I used the six attributes to fit the OLS Regression model with SALES_QTY as dependent variables. And I also tried two ways of regression: with intercept, without intercept.
Model results:

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:              SALES_QTY   R-squared (uncentered):                   0.344
Model:                            OLS   Adj. R-squared (uncentered):              0.333
Method:                 Least Squares   F-statistic:                              30.20
Date:                Sun, 12 Jan 2020   Prob (F-statistic):                    4.34e-29
Time:                        18:35:04   Log-Likelihood:                         -2267.9
No. Observations:                 351   AIC:                                      4548.
Df Residuals:                     345   BIC:                                      4571.
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
IF_GENERAL_PRICE               30.9727     16.276      1.903      0.058      -1.039      62.985
SEASON_SINCE_1ST_SKU_LAUNCH    11.6486      2.793      4.170      0.000       6.155      17.143
IF_RETRO                      348.0873     57.106      6.095      0.000     235.768     460.407
IF_SEASONAL_SILO_SU           -84.9327     40.376     -2.104      0.036    -164.346      -5.519
NUMSKU_SAME_STYLE_IN1SSN       24.6858      4.023      6.137      0.000      16.774      32.598
OMD_DAYS_SINCE_SEASON_BEGIN    -0.4940      0.379     -1.305      0.193      -1.239       0.251
==============================================================================
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SALES_QTY   R-squared:                       0.147
Model:                            OLS   Adj. R-squared:                  0.132
Method:                 Least Squares   F-statistic:                     9.880
Date:                Sun, 12 Jan 2020   Prob (F-statistic):           4.62e-10
Time:                        18:36:00   Log-Likelihood:                -2267.7
No. Observations:                 351   AIC:                             4549.
Df Residuals:                     344   BIC:                             4576.
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                          16.5777     22.804      0.727      0.468     -28.275      61.431
IF_GENERAL_PRICE               26.9924     17.182      1.571      0.117      -6.803      60.788
SEASON_SINCE_1ST_SKU_LAUNCH    11.1057      2.893      3.839      0.000       5.415      16.796
IF_RETRO                      344.5794     57.348      6.009      0.000     231.782     457.377
IF_SEASONAL_SILO_SU           -80.3551     40.891     -1.965      0.050    -160.783       0.073
NUMSKU_SAME_STYLE_IN1SSN       20.9387      6.540      3.202      0.001       8.075      33.802
OMD_DAYS_SINCE_SEASON_BEGIN    -0.5872      0.400     -1.468      0.143      -1.374       0.200
==============================================================================
Omnibus:                      295.932   Durbin-Watson:                   1.962
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5600.451
Skew:                           3.501   Prob(JB):                         0.00
Kurtosis:                      21.273   Cond. No.                         196.
==============================================================================
```

According to two models' results, R-squared decreased, this is understandable, because we deleted many attributes. And two variables are not significant anymore, I found four significant attributes:

'SEASON_SINCE_1ST_SKU_LAUNCH',
'IF_RETRO',
'IF_SEASONAL_SILO_SU',
'NUMSKU_SAME_STYLE_IN1SSN',

4. Second part of attributes regression

I used the four attributes to fit the OLS Regression model with SALES_QTY as dependent variables. And I also tried two ways of regression: with intercept, without intercept.

Model results:

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:              SALES_QTY   R-squared (uncentered):                   0.335
Model:                            OLS   Adj. R-squared (uncentered):              0.327
Method:                 Least Squares   F-statistic:                              43.71
Date:                Sun, 12 Jan 2020   Prob (F-statistic):                    1.06e-29
Time:                        18:41:06   Log-Likelihood:                         -2270.4
No. Observations:                 351   AIC:                                      4549.
Df Residuals:                     347   BIC:                                      4564.
Df Model:                           4
Covariance Type:            nonrobust
===============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
SEASON_SINCE_1ST_SKU_LAUNCH   12.3851      2.781      4.454      0.000       6.916      17.854
IF_RETRO                     318.0802     52.474      6.062      0.000     214.874     421.286
IF_SEASONAL_SILO_SU         -101.8881     38.934     -2.617      0.009    -178.464     -25.312
NUMSKU_SAME_STYLE_IN1SSN      26.7342      2.910      9.188      0.000      21.011      32.457
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SALES_QTY   R-squared:                       0.135
Model:                            OLS   Adj. R-squared:                  0.125
Method:                 Least Squares   F-statistic:                     13.54
Date:                Sun, 12 Jan 2020   Prob (F-statistic):           2.89e-10
Time:                        18:42:20   Log-Likelihood:                -2270.0
No. Observations:                 351   AIC:                             4550.
Df Residuals:                     346   BIC:                             4569.
Df Model:                           4
Covariance Type:            nonrobust
===============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                         17.2540     20.463      0.843      0.400     -22.994      57.502
SEASON_SINCE_1ST_SKU_LAUNCH   11.8525      2.853      4.155      0.000       6.242      17.463
IF_RETRO                     308.5250     53.705      5.745      0.000     202.896     414.154
IF_SEASONAL_SILO_SU          -93.4742     40.208     -2.325      0.021    -172.557     -14.392
NUMSKU_SAME_STYLE_IN1SSN      21.7849      6.552      3.325      0.001       8.898      34.672
==============================================================================
Omnibus:                      296.491   Durbin-Watson:                   1.953
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5507.118
Skew:                           3.521   Prob(JB):                         0.00
Kurtosis:                      21.082   Cond. No.                         22.7
==============================================================================
```

Now, all four attributes stayed statistically significant, we can trust they can affect SALES_QTY.

Interpretation

Due to the intercept is not significant, we use the result from the model without intercept.
Let's check out the correlation coefficients of these four attributes and SALES_QTY:

	Ols coef	SALES_QTY coef
SEASON_SINCE_1ST_SKU_LAUNCH	12.39	0.2268
IF_RETRO	318.08	0.2234
IF_SEASONAL_SILO_SU	-101.89	-0.0795
NUMSKU_SAME_STYLE_IN1SSN	26.73	0.0851

The first two attributes have high correlation coefficients with SALES_QTY, but the last two attributes have relatively low correlation coefficients. Therefore, we cannot say attributes with low correlation coefficients must be insignificant.

Except for these four attributes, we cannot identify how other product attributes affect product sales.

'Number of seasons since the SKU was launched' has positive relation with sales unit, more specifically, if one more season passed since the SKU was launched, around 12 more units of product will be sold. This is very reasonable. Because total sales unit must be increasing with time passing.

'If it is a Retro product' has positive relation with sales unit, more specifically, if it is a Retro product, around 318 more units of product will be sold. This is also understandable, because a Retro product must be popular, otherwise Nike would be less likely to reproduce it. And vintage style is pretty popular for recent years.

'If the product is Summer special' has negative relation with sales unit, more specifically, if it is a summer special product, around 102 less units of product will be sold. This maybe because that summer special products are less useful than all-season product, people have lower level of demand.

'Number of products with the same style' has positive relation with sales unit, more specifically, if number of products with the same style increase by 1, around 27 more units of product will be sold. This is also understandable, because the higher number of products with the same style, the more popular the style is.