**Multiple Linear Regression Report**

# **1. Introduction**

The purpose of this analysis was to verify how much two search advertising strategies (organic search and paid search/adwords keywords) and company assets influence the achievement of revenue and profit. In other words, we want to examine how much of profit and revenue can be determined using these three factors.

All information related to search advertising strategies were obtained from ‘weboutlook.com’ website (the full URL was defined by concatenation between the actual domain of the company with this website) and the asset was from Fortune 1000 list. The organic search and paid search were obtained from ‘Organic Traffic’ and ‘Adwords Traffic’ fields, respectively, while the company’s asset was from ‘Assets ($m)’ field.

# **2. Exploratory Data Analysis**

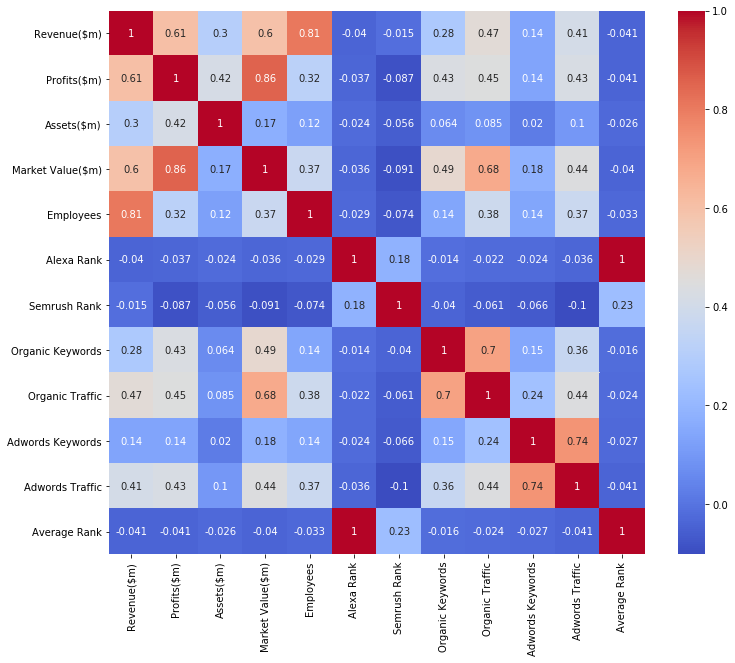
As we wanted to verify the influence of the factors in two outcomes (revenue and profit), the analysis was divided in two independent process, although the data source was just one.

For each process was evaluated 3 (three) models: Ordinary Least Squares (OLS)/Linear Regression, Ridge Regression, and Lasso Regression.

## **2.1. Correlations Analysis**

Using the seaborn heatmap was possible to see how the variables are correlated, remembering that our outcomes are ‘revenue’ and ‘profit’.

Analysing the graph, we can see that, in general, the variables are poorly correlated. The three highest correlation observed are 0.86 (between ‘revenue’ and ‘market value’), 0.81 (between ‘revenue’ and ‘employees’), and 0.74 (between ‘adwords traffic’ and ‘adwords keywords’). The biggest part of the graph shows correlations below than 0.5. We must ignore the ‘average rank’ variable because this variable is the sum of ‘alexa rank’ and ‘semrush rank’ variables. Another interesting point is in some cases there are negative correlations, however, all values are very close to zero.



**Figure 1 – Correlations heatmap**

Considering only the variables that will be used in modeling process, the correlation values are:

TABLE 1 – Correlations among outcomes and selected variables to model

|  |  |  |  |
| --- | --- | --- | --- |
| Outcome | Assets | Organic Traffic | Adwords Traffic |
| Revenue | 0.30 | 0.47 | 0.41 |
| Profit | 0.42 | 0.45 | 0.43 |

## **2.2. Histogram Analysis**

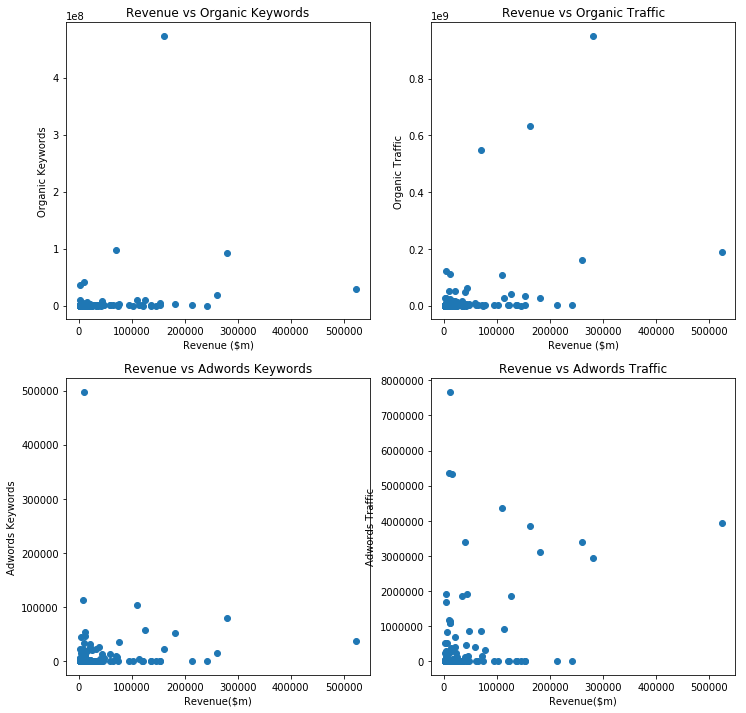
The sequence of charts shows the histograms before the data transformation and we can see that the data are very right skewed.

|  |  |
| --- | --- |
|  |  |

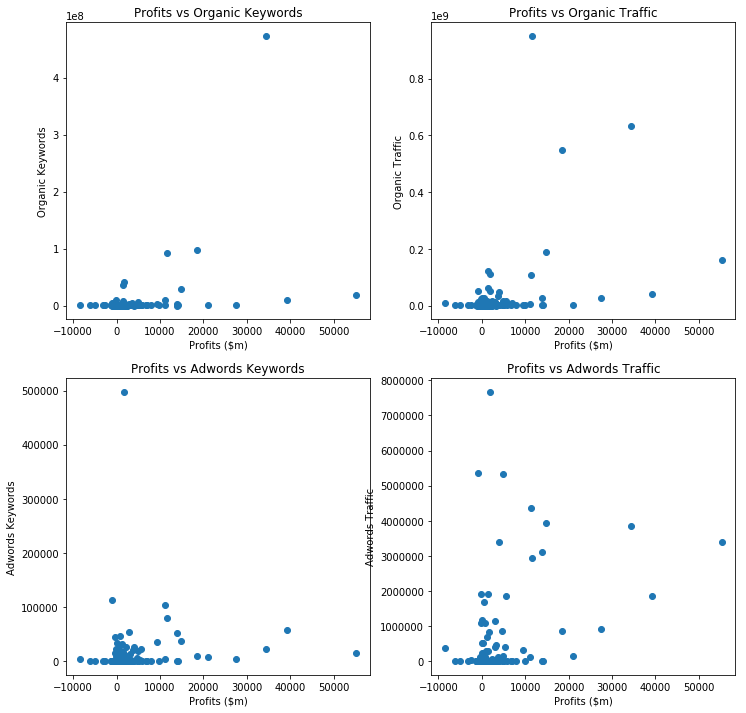
**Figure 2 – Histograms of the variables**

## **2.3. Scatter plot Analysis**

Regarding the scatterplots, also before the data transformation, the data are very clustered with a few points out (some points could be considered outliers because they live very far of the cluster). The first set of scatter plots shows the



**Figure 3 – Scatter plot with the ‘revenue’ outcome**



**Figure 4 – Scatter plot with the ‘profit’ outcome**

# **3. Data Modeling**

To perform all modeling, it was used the ‘*random-state = 0*’ to guarantee the reproducibility.

## **3.1. Analysis considering the outcome ‘Revenue’**

### *3.1.1. Ordinary Least Squares Model Assessment*

From the statistical results of model adjustment, we can conclude that both model and coefficients are statistically significant, because the F-statistic of de model was 35.29 (p-value < 7.59e-18) and all p-values of each coefficients were less than 0.002 (t-test greater than 3).

The model can explain only 36.9% of the variation in revenue (adjusted R-squared = 0.369). Although the explanation is low, all variables are significative for the model, what is a good information to try to predict the revenue from these variables. The results are presented below.

OLS Regression Results

==============================================================================

Dep. Variable: Revenue($m) R-squared: 0.380

Model: OLS Adj. R-squared: 0.369

Method: Least Squares F-statistic: 35.29

Date: Sun, 14 Jun 2020 Prob (F-statistic): 7.59e-18

Time: 16:16:57 Log-Likelihood: 179.85

No. Observations: 177 AIC: -351.7

Df Residuals: 173 BIC: -339.0

Df Model: 3

Covariance Type: nonrobust

===================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------

const 0.0219 0.007 3.129 0.002 0.008 0.036

Organic Traffic 0.3613 0.089 4.059 0.000 0.186 0.537

Adwords Traffic 0.3874 0.086 4.526 0.000 0.218 0.556

Assets($m) 0.2971 0.086 3.444 0.001 0.127 0.467

==============================================================================

Omnibus: 187.251 Durbin-Watson: 2.087

Prob(Omnibus): 0.000 Jarque-Bera (JB): 5319.410

Skew: 4.012 Prob(JB): 0.00

Kurtosis: 28.630 Cond. No. 16.5

==============================================================================

### *3.1.2. Linear Regression, Ridge Regression, and Lasso Regression*

For ridge and lasso regression were employed cross-validation technique to determine the best or optimal value of alpha parameter (α). It is interesting to note the relationships among OLS, ridge and lasso regression show in the equations below. Actually, ridge and lasso work as a regulator adjusting OLS regression.

The table below presents the best alpha values for Ridge and Lasso regressions as well as Root Mean Squared Error (RMSE), statistic used to evaluate the performance of the three predictive models. In this case, the Ridge Regression was the best model because it has the lowest RMSE. These values showed are referenced to transformed data (min-max transformation).

TABLE 2 – Performance of the regression model for ‘revenue’ outcome

|  |  |  |
| --- | --- | --- |
| Regression Method | Alpha | RMSE |
| OLS | - | 0.068156 |
| Ridge | 1.0 | 0.060490 |
| Lasso | 0.01 | 0.063866 |

## **3.2. Analysis considering the outcome ‘Profit’**

### *3.2.1. Ordinary Least Squares Model Assessment*

Similarly the first outcome, the statistical results of model adjustment shows that both model and coefficients are statistically significant, because the F-statistic of de model was 43.50 (p-value < 5.32e-21) and all p-values of each coefficients were less than 0.001 (t-test greater than 3.5).

The model can explain 42.0% of the variation in profit (adjusted R-squared = 0.420). Although the explanation is almost the half, all variables are significative for the model, what is a good information to try to predict the profit from these variables. The results are presented below.

OLS Regression Results

==============================================================================

Dep. Variable: Profits($m) R-squared: 0.430

Model: OLS Adj. R-squared: 0.420

Method: Least Squares F-statistic: 43.50

Date: Sun, 14 Jun 2020 Prob (F-statistic): 5.32e-21

Time: 16:16:57 Log-Likelihood: 237.85

No. Observations: 177 AIC: -467.7

Df Residuals: 173 BIC: -455.0

Df Model: 3

Covariance Type: nonrobust

===================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------

const 0.1432 0.005 28.397 0.000 0.133 0.153

Organic Traffic 0.2253 0.064 3.512 0.001 0.099 0.352

Adwords Traffic 0.3370 0.062 5.464 0.000 0.215 0.459

Assets($m) 0.3074 0.062 4.946 0.000 0.185 0.430

==============================================================================

Omnibus: 245.201 Durbin-Watson: 2.007

Prob(Omnibus): 0.000 Jarque-Bera (JB): 28167.214

Skew: 5.569 Prob(JB): 0.00

Kurtosis: 63.788 Cond. No. 16.5

==============================================================================

### *3.2.2. Linear Regression, Ridge Regression, and Lasso Regression*

In similar way, the table below presents the best alpha values for Ridge and Lasso regressions as well as Root Mean Squared Error (RMSE) obtained for ‘profit’ outcome. In this case, the OLS Regression was the best model because it has the lowest RMSE. These values showed are referenced to transformed data (min-max transformation).

TABLE 3 – Performance of the regression model for ‘profit’ outcome

|  |  |  |
| --- | --- | --- |
| Regression Method | Alpha | RMSE |
| OLS | - | 0.074578 |
| Ridge | 1.0 | 0.075064 |
| Lasso | 0.01 | 0.086892 |

# **4. Making Predictions**

After the model development phase, it was performed a prediction using those four models defined in the previous section. To predict both revenue and profit forecasted are shown in the tables below, remembering that for Revenue was used Ridge Regression and for Profit was OLS Regression due to RMSE minimum.

TABLE 4 – Revenue Forecasted by Ridge Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Asset | Organic Traffic | Adwords Traffic | Revenue |
| Transformed Value | 0.25 | 0.15 | 0.21 | 0.166738 |
| Actual Value | $ 876,508m | 142,562,598 | 1,611,306 | $ 89,030.6m |

TABLE 5 – Profit Forecasted by OLS Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Asset | Organic Traffic | Adwords Traffic | Profit |
| Transformed Value | 0.25 | 0.15 | 0.21 | 0.324591 |
| Actual Value | $ 876,508m | 142,562,598 | 1,611,306 | $ 12,190.6m |

In short, based on the dataset and the adjusted models (random\_state = 0), it is predicted a **Revenue** about **$89mi** and a **Profit** around **$12.2mi**.

# **5. Taking Decisions**

As we can see in the above tables, in general, the search advertising strategies work very well because in both cases and all tested models the presented results were expressive. In short, we can affirm that the companies should use these strategies to try to stay or become more competitive in its the business environment.