MKT 3019 Data Analysis Project: New-Ark Shoes Ltd

Student Number: 190685094

Submitted Date: 18th May 2022

Program Code: NN52

Total Word Count: 2192 Excluding Titles Sub-titles, Captions, References, Tables and Figures, Appendix, Table of Contents, Table of Figures, Table of Tables

Table of Contents

Table of Tables

Table of Figures

1. Predictive business intelligence.

1.1 Linear Regression Model.

1.1.1 Introduction of key indicators.

For this report, linear regression will be used for the initial analysis, it refers to the arrangement of data sets into 2 categories, dependent and independent variables. In this case, the multivariable linear regression will be specifically used in this investigation. It is a subdivision of linear regression which associates one target variable to multiple predictors and can be summarized into y =b0 +b1 x1 +b2 x2 +...+bp xp +ϵ where x1 and x2 are different predictors and y is the target variable.

Several key indicators are needed to understand the quality of the model their acceptable range is displayed in the table below. Firstly, the Pearson correlation coefficient which represents the strength of correlation between variables. Secondly, Adjusted r-square are the key benchmarks for evaluating the quality of the representation, in other words, they are the percentage of data scattering that can be covered by the linear model. Lastly, the P-value are the indicator of the significance possessed by on predictor variable within the model. Usually, the smaller the more significant in the model.

Benchmark	Stong	Moderate	Poor
Pearson Correlation Value	>=0.7	0.4<>0.7	<=0.4
(Adjusted) R-Squared		1<>0.3	<0.3
P-Value	<0.05		>0.05

Table 1—1 Benchmark Acceptable Range

1.1.2 Linear regression model #1.

Upon importing the data into the software, it will be divided into “Estimation” and “Validation” groups. The first is used for model generation and the second is used to check the validity of the

model. After generating the linear model, an assessment of the model and its predictors will be made. Next, the “Stepwise tool” is used to refine the model by eliminating the poorly performed predictors, this is followed by the “Nested tool” which compares the two models and determines whether significant improvement has been made. Lastly, the “Association Analysis” will be used to identify the strength of each individual pair of targets and predictors (Appendix 6.4).

Report for Linear Model Linear_Regression_Sales_Per_order

Basic Summary

Call:
lm(formula = Sales.... ~ Household.Type + Category + Quantity + Discount + Profit.... + Number.of.Web.Visit.s..Before.Purchase + Time.Spent.on.the.Website.Before.Purchase..Mins. + Number.of.Pages.Visited.Before.Purchase, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-684.3	-134.1	-36.1	38.6	4292.9

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	129.860	44.66293	2.9076	0.0037 **
Household.TypeFamily	23.023	19.86404	1.1590	0.24666
Household.TypeSingles	-11.647	25.77929	-0.4518	0.65149
CategoryMen's Shoe	-58.525	31.52276	-1.8566	0.06359 .
CategoryWomen's Shoe	-258.374	23.60761	-10.9445	< 2.2e-16 ***
Quantity	44.685	4.20289	10.6319	< 2.2e-16 ***
Discount	198.849	64.14468	3.1000	0.00198 **
Profit....	2.214	0.05363	41.2833	< 2.2e-16 ***
Number.of.Web.Visit.s..Before.Purchase	1.433	6.11857	0.2342	0.81487
Time.Spent.on.the.Website.Before.Purchase..Mins.	1.715	0.74508	2.3020	0.02149 *
Number.of.Pages.Visited.Before.Purchase	-3.123	2.14094	-1.4586	0.14491

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 314.92 on 1295 degrees of freedom
Multiple R-squared: 0.6904, Adjusted R-Squared: 0.688

Table 1—2 Result of Linear Regression Model #1

Report for Linear Model Sales_Per_Order_Step_Wise

Basic Summary

Call:
lm(formula = Sales.... ~ Category + Quantity + Discount + Profit.... + Time.Spent.on.the.Website.Before.Purchase..Mins. + Number.of.Pages.Visited.Before.Purchase, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-689.2	-136.0	-33.2	36.2	4306.4

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	143.808	37.59819	3.825	0.00014 ***
CategoryMen's Shoe	-58.621	31.50440	-1.861	0.06301 .
CategoryWomen's Shoe	-258.246	23.60184	-10.942	< 2.2e-16 ***
Quantity	44.542	4.19846	10.609	< 2.2e-16 ***
Discount	200.957	64.11611	3.134	0.00176 **
Profit....	2.214	0.05358	41.314	< 2.2e-16 ***
Time.Spent.on.the.Website.Before.Purchase..Mins.	1.701	0.74360	2.288	0.02232 *
Number.of.Pages.Visited.Before.Purchase	-3.062	2.13011	-1.438	0.1508

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 314.89 on 1298 degrees of freedom
Multiple R-squared: 0.6898, Adjusted R-Squared: 0.6881

Table 1—3 Regression Result Model #1 Stepwise

The Effect of Removing the Variables Household.Type and Number.of.Web.Visit.s..Before.Purchase from Linear_Regression_Sales_Per_order

DF	Sum of Squares	F	Pr(>F)
3	270920.85	0.9106	0.43518

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1—4 Linear Model #1 Nested Test Result

ayout

Pearson Correlation Analysis

Full Correlation Matrix

	Sales....	Discount	Number.of.Web.Visit.s..Before.Purchase	Time.Spent.on.the.Website.Before.Purchase..Mins.	Number.of
Sales....	1.0000000	-0.1142192	0.0312498	-0.0187230	
Discount	-0.1142192	1.0000000	0.0012830	0.0054369	
Number.of.Web.Visit.s..Before.Purchase	0.0312498	0.0012830	1.0000000	0.0161189	
Time.Spent.on.the.Website.Before.Purchase..Mins.	-0.0187230	0.0054369	0.0161189	1.0000000	
Number.of.Pages.Visited.Before.Purchase	0.0069840	0.0069795	0.0043578	-0.0075221	

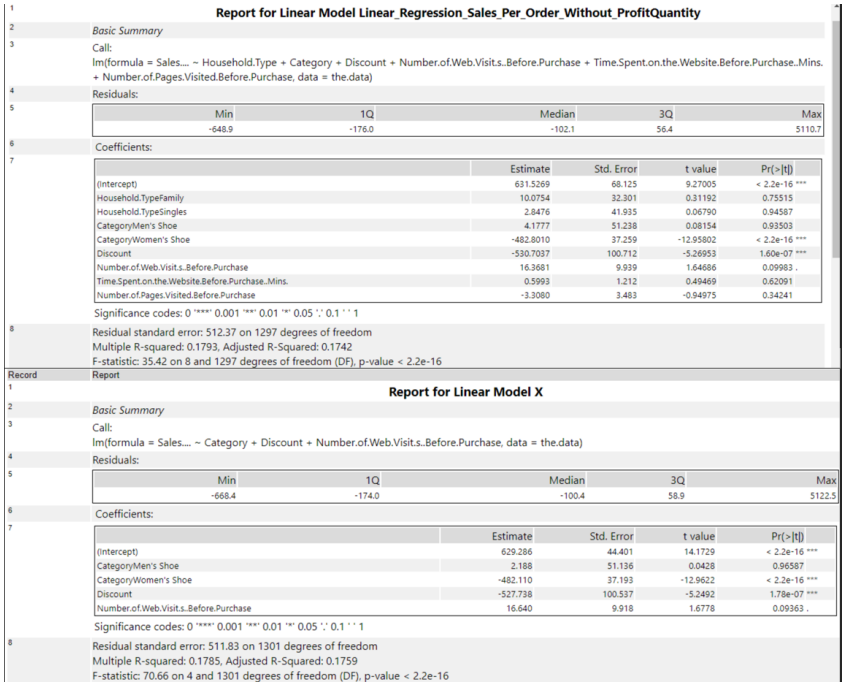
Matrix of Corresponding p-values

	Sales....	Discount	Number.of.Web.Visit.s..Before.Purchase	Time.Spent.on.the.Website.Before.Purchase..Mins.	Number.of
Sales....		3.7054e-06	2.0689e-01	4.4960e-01	
Discount	3.7054e-06		9.5868e-01	8.2623e-01	
Number.of.Web.Visit.s..Before.Purchase	2.0689e-01	9.5868e-01		5.1510e-01	
Time.Spent.on.the.Website.Before.Purchase..Mins.	4.4960e-01	8.2623e-01	5.1510e-01		
Number.of.Pages.Visited.Before.Purchase	7.7793e-01	7.7807e-01	8.6032e-01	7.6133e-01	

Table 1—5 Pearson Correlation Matrix

The initial result is displayed in the first figure, the overall validity of the model is acceptable as the r-square value is 0.688. Removing the predictors with low significance yielded a new model with a similar lar r-value of 0.688 and both models proved to be not significantly different from one another. Nevertheless, the problem with it is that model is Pearson Correlation Coefficient of “discount” is negative (Table 1-5) while the Regression Coefficient (Estimate) is positive, so they indicate 2 different trend directions which are contradictory. However, such phenomena are common in multivariable linear regression, and it is called Positive Net Suppression. It occurs when 2 or more predictors in the model are so strong that they suppressed weaker predictors and cause them to have a certain degree of variation in the regression (Nickerson, 2008). In more practical senses, quantity and profit are more directly linked with sales which inherently make the relation stronger, thus causing the regression to be superficially “accurate” while overshadowing other predictors. What’s more. Profit as a predictor of Sales is not a logically sound relation because the former does not “cause” the latte

1.1.3 Linear regression model #2.



Report for Linear Model Linear_Regression_Sales_Per_Order_Without_ProfitQuantity

Basic Summary

Call:
lm(formula = Sales.... ~ Household.Type + Category + Discount + Number.of.Web.Visit.s..Before.Purchase + Time.Spent.on.the.Website.Before.Purchase..Mins. + Number.of.Pages.Visited.Before.Purchase, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-648.9	-176.0	-102.1	56.4	5110.7

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	631.5269	68.125	9.27005	< 2.2e-16 ***
Household.TypeFamily	10.0754	32.301	0.31192	0.75515
Household.TypeSingles	2.8476	41.935	0.06790	0.94587
CategoryMen's Shoe	4.1777	51.238	0.08154	0.93503
CategoryWomen's Shoe	-482.8010	37.259	-12.95802	< 2.2e-16 ***
Discount	-530.7037	100.712	-5.26953	1.60e-07 ***
Number.of.Web.Visit.s..Before.Purchase	16.3681	9.939	1.64686	0.09983 .
Time.Spent.on.the.Website.Before.Purchase..Mins.	0.5993	1.212	0.49469	0.62091
Number.of.Pages.Visited.Before.Purchase	-3.3080	3.483	-0.94975	0.34241

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 512.37 on 1297 degrees of freedom
Multiple R-squared: 0.1793, Adjusted R-Squared: 0.1742
F-statistic: 35.42 on 8 and 1297 degrees of freedom (DF), p-value < 2.2e-16

Record | Report

Report for Linear Model X

Basic Summary

Call:
lm(formula = Sales.... ~ Category + Discount + Number.of.Web.Visit.s..Before.Purchase, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-668.4	-174.0	-100.4	58.9	5122.5

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	629.286	44.401	14.1729	< 2.2e-16 ***
CategoryMen's Shoe	2.188	51.136	0.0428	0.96587
CategoryWomen's Shoe	-482.110	37.193	-12.9622	< 2.2e-16 ***
Discount	-527.738	100.537	-5.2492	1.78e-07 ***
Number.of.Web.Visit.s..Before.Purchase	16.640	9.918	1.6778	0.09363 .

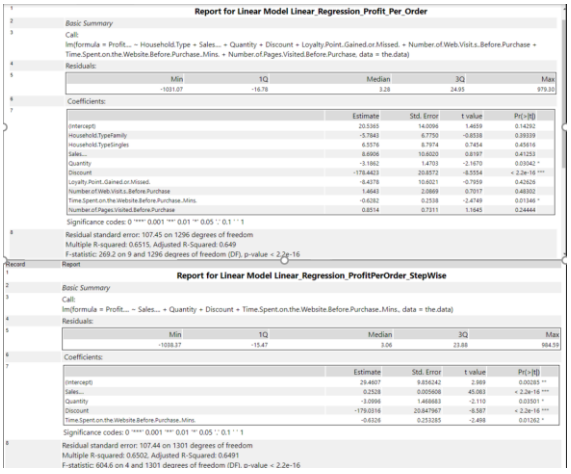
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 511.83 on 1301 degrees of freedom
Multiple R-squared: 0.1785, Adjusted R-Squared: 0.1759
F-statistic: 70.66 on 4 and 1301 degrees of freedom (DF), p-value < 2.2e-16

Table 1—6 Regression Result Model #2 + Stepwise

To improve the result, quantity and profit are removed. However, r -the square decreases dramatically to 0.174 (Table 1-6) which is way below the acceptive level. Removing the underperforming predictors still resulted in the r-square remaining at an appalling level of 0.1759. In summary, none of these models can be considered useful and shall be disregarded.

1.1.4 Linear regression model #3.



Report for Linear Model Linear_Regression_Profit_Per_Order

Basic Summary

Call:
lm(formula = Profit.... ~ Household.Type + Sales.... + Quantity + Discount + Loyalty.Point..Gained.or.Missed. + Number.of.Web.Visit.s..Before.Purchase + Time.Spent.on.the.Website.Before.Purchase..Mins. + Number.of.Pages.Visited.Before.Purchase, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-1031.07	-16.76	3.28	24.95	979.30

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.5365	14.0096	1.4659	0.14282
Household.TypeFamily	-5.7843	6.7750	-0.8538	0.39339
Household.TypeSingles	6.5576	8.7974	0.7454	0.45616
Sales....	8.6906	10.6620	0.8197	0.41253
Quantity	-3.1862	1.4703	-2.1670	0.03042 *
Discount	-178.4403	20.8572	-8.5554	< 2.2e-16 ***
Loyalty.Point..Gained.or.Missed.	-6.4778	10.6021	-0.7959	0.42626
Number.of.Web.Visit.s..Before.Purchase	1.4643	2.0869	0.7017	0.48302
Time.Spent.on.the.Website.Before.Purchase..Mins.	-0.6282	0.2538	-2.4749	0.01346 *
Number.of.Pages.Visited.Before.Purchase	0.8514	0.7311	1.1645	0.24444

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 107.45 on 1296 degrees of freedom
Multiple R-squared: 0.6515, Adjusted R-Squared: 0.649
F-statistic: 269.2 on 9 and 1296 degrees of freedom (DF), p-value < 2.2e-16

Record | Report

Report for Linear Model Linear_Regression_ProfitPerOrder_StepWise

Basic Summary

Call:
lm(formula = Profit.... ~ Sales.... + Quantity + Discount + Time.Spent.on.the.Website.Before.Purchase..Mins., data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-1038.37	-15.47	3.06	23.88	984.59

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	29.4607	9.856242	2.989	0.00285 **
Sales....	0.2528	0.005608	45.083	< 2.2e-16 ***
Quantity	-3.0996	1.466683	-2.110	0.03501 *
Discount	-179.0316	20.847067	-8.587	< 2.2e-16 ***
Time.Spent.on.the.Website.Before.Purchase..Mins.	-0.6326	0.253285	-2.498	0.01262 *

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 107.44 on 1301 degrees of freedom
Multiple R-squared: 0.6502, Adjusted R-Squared: 0.6491
F-statistic: 604.6 on 4 and 1301 degrees of freedom (DF), p-value < 2.2e-16

Table 1—7 Regression Result Model # 3 + Stepwise

Record | Layout

1

The Effect of Removing the Variables Household.Type, Loyalty.Point..Gained.or.Missed., Number.of.Web.Visit.s..Before.Purchase and Number.of.Pages.Visited.Before.Purchase from Linear_Regression_Profit_Per_Order

DF	Sum of Squares	F	Pr(>F)
5	54337.38	0.9412	0.45318

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1—8 Nested Test for Linear Regression Model #3

Since setting the target variable as Sales is unlikely to produce a good model, then shifting the target to Profit might be a viable option. As such, the previous predictors will be included along with quantity and sales. The initial result shows a prominent r-square value of 0.649 (Table 1-7) which is a vast improvement over the previous ones. Plus, the Regression Coefficient agrees with Pearson Coefficient which means no Positive Net Suppression effect appeared. Stepwise resulted in a model with a similar r-square which is proven to be not significantly different from the first one. For discount and quantity, they formed a negative slope in relation to profit. This means the more discounts applied per order, the lower profit is generated in that order. Alarmingly, quantity also displays such a trend, which means the more item purchased in each order the less profit it generates. This indicates a potential pricing scheme deficiency, as more items purchased are not effectively converted to profit due over-aggressive discount strategy (demonstrated by the more extreme downward slope)

1.1.5 Linear regression model #4.

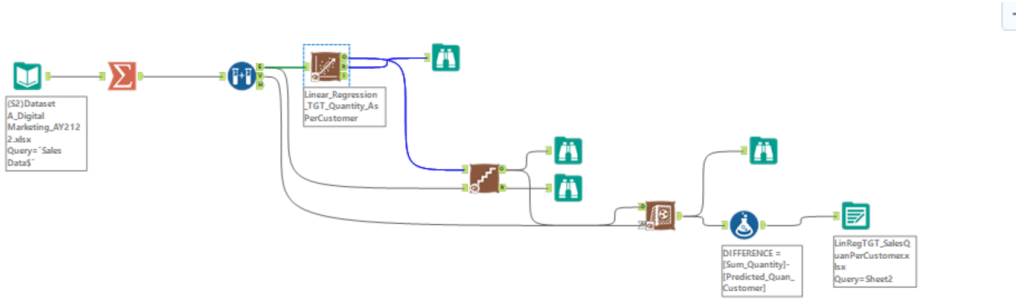


Figure 1—1 Procedure for Linear Regression Model #4

Min	1Q	Median	3Q	Max
-11.453	-2.348	-0.293	1.826	17.997

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.49254	0.351167	1.403	0.16159	
Sum_Discount	1.05794	0.541818	1.953	0.05163	.
Sum_Number.of.Web.Visit.s..Before.Purchase	0.37783	0.066541	5.678	2.78e-08	***
Sum_Time.Spent.on.the.Website.Before.Purchase..Mins.	0.04471	0.007505	5.958	6.01e-09	***
Sum_Number.of.Pages.Visited.Before.Purchase	0.13343	0.024803	5.380	1.33e-07	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.0351 on 365 degrees of freedom
Multiple R-squared: 0.815, Adjusted R-Squared: 0.813
F-statistic: 402 on 4 and 365 degrees of freedom (DF), p-value < 2.2e-16

Table 1—9 Regression Result Model # 4

Network related predictors never presented a strong correlation in previous models, a new model should be used to better represent them. To do that, a new approach must use which sums up every numerical predictor under each distinctive customer. Thus, creating aggerated values which represent the customer behaviour over the course of time spam. Then, Sum-Quantity will be selected as the target variable and its predictor includes Summation of Discount and another web-related predictor. The initial result is displayed above, and the model presents a much superior strength with an r-square value of 0.813 which means over 81.3 % of the target variable can be explained by the predictors. Even Stepwise could not further improve the r-square. Moreover, every web-related predictor shows a strong sign of significance in the model as their p-values are very close to 0 which means they are very likely to be significant in the model. In a particle sense, because of the positive slope of 0.377 and 0.133, the more website and pages a customer visit before making a purchase over time, the more likely he orders more items in that time period.

1.2 Logistic Regression

1.2.1 Introduction

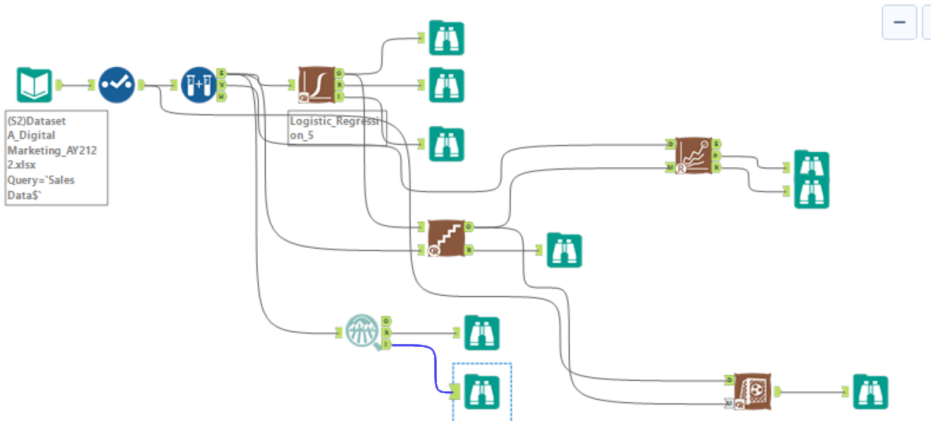


Figure 1—3 Procedure of Logistic Regression #1

When it comes to the perditions of the binary variables, or in other words, variables with only 2 outcomes, it is best to produce the predictive model with logistic regression. Since advertisement and marketing campaign is an important part of the business operation. Therefore, investigating what quality a customer should possess at the time of purchase so that they respond to the email campaign. The process works on a similar flow, but models are compared by the “Model Comparison tooland the quality of the model is presented by “McFadden R-Square”.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.17547	0.2545	0.6896	0.49047	
Household.TypeFamily	0.23343	0.1378	1.6943	0.09021	.
Household.TypeSingles	0.08223	0.1757	0.4679	0.63984	
RegionEast	-0.42524	0.2550	-1.6675	0.09541	.
RegionMidlands	0.10852	0.1924	0.5639	0.57281	
RegionNorth	-0.04397	0.2064	-0.2130	0.8313	
RegionNorth-East	0.14247	0.3030	0.4702	0.63822	
RegionNorth-West	0.02349	0.2272	0.1034	0.91764	
RegionSouth	0.15385	0.3255	0.4726	0.63647	
RegionSouth-East	-0.24663	0.2494	-0.9888	0.32274	
RegionSouth-West	0.27974	0.2980	0.9386	0.34793	
RegionWest	-0.25338	0.5009	-0.5059	0.61295	
Mode.of.PaymentCredit Card	0.18788	0.1910	0.9835	0.32534	
Mode.of.PaymentDebit Card	0.16930	0.1965	0.8615	0.38895	
Mode.of.PaymentGift Card	0.14074	0.1940	0.7256	0.46808	
Mode.of.PaymentPaypal	-0.10062	0.1911	-0.5266	0.59845	
Loyalty.Card.Holder..Yes	-0.14706	0.1210	-1.2157	0.22411	
Downloaded.Discount.Voucher.for.Future.Purchase.Yes	-0.14719	0.1207	-1.2190	0.22283	
Type.s..of.Catalogue.Sent.before.PurchaseOnline	-0.33655	0.1497	-2.2476	0.0246	*
Type.s..of.Catalogue.Sent.before.PurchasePhysical (Paper Based)	-0.18654	0.1507	-1.2376	0.21586	

Null deviance: 1584.1 on 1143 degrees of freedom
Residual deviance: 1562.2 on 1122 degrees of freedom
McFadden R-Squared: 0.01383, Akaike Information Criterion 1604

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.0000758	0.16322	-0.0004644	0.99963	
Number.of.Web.Visit.s..Before.Purchase	0.0703843	0.04125	1.7063502	0.08794	.
Type.s..of.Catalogue.Sent.before.PurchaseOnline	-0.3187919	0.14779	-2.1571153	0.031	*
Type.s..of.Catalogue.Sent.before.PurchasePhysical (Paper Based)	-0.1641876	0.14855	-1.1053009	0.26903	

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial taken to be 1)

Null deviance: 1584.1 on 1142 degrees of freedom
Residual deviance: 1576.5 on 1139 degrees of freedom
McFadden R-Squared: 0.004833, Akaike Information Criterion 1584

Table 1—10 Logistic Regression Result Model #1 +Stepwise

As the result shown above, the r-square shows an appalling value of 0.01383 which means only 1.3% target variable can be explained by the predictors and is too low of the regression to represent the actual data set. Similar conditions can be observed in the strength of the predictors, as none of them is very effective in predicting the target variable.

Record	Layout
1	**Model Comparison Report**
2	**Fit and error measures**

Model	Accuracy	Accuracy_No	Accuracy_Yes	F1	AUC
Logistic_Regression_TGT_EmailResponds_StepWise	0.4959	0.3279	0.6667	0.3961	0.4976

Table 1—11 Model Comparison Test Result

Even worse, when using the Stepwise Tool to trim the least probable predictors into a new model, no significant improvement to the model has been observed. According to the Figure, the model is only 49.59% accurate in predicting the outcome. Another model is based on the classification model which segmented quantitate data into different levels. However, the result is not ideal either as indicated in Appendix 6.3.

Part 2 Digital Marketing; KPI Identifying

2.1 KPI: Most prominent keyword contender

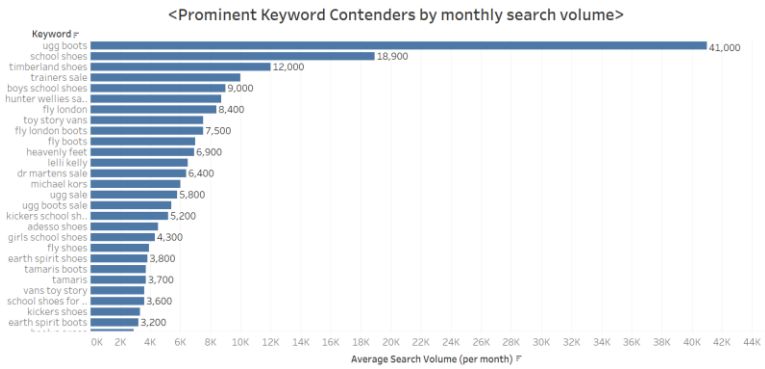


Figure 2—1 Sorted Bar Chart of Top Keyword Searches per Month

Keywords are terms which customers put in the search engines to reach the desired information. After the input, the search results are displayed on the result page, on which they may find the company website link and proceed to access if they find the information is relevant. The figure above indicated the average search volume per month of each keyword, it represents the number of instances of people who input a particular keyword into a search engine. Since the website display at the top of the result lists is often highly related to the keyword, then modifying the company's website to be more relevant to the popular keywords may increase the exposure greatly. Thus improving the number of visits conversion rate. In this case, as indicated in Figure. The recurring theme of the top keywords often related to kid shoes and women's shoes, in particular, boots and school shoes are the most popular. Therefore, the marketing on the website should emphasise these terms by adding relevant promotions and discontent so that the website traffic may improve.

2.2 KPI Composition of device used for visiting the website.

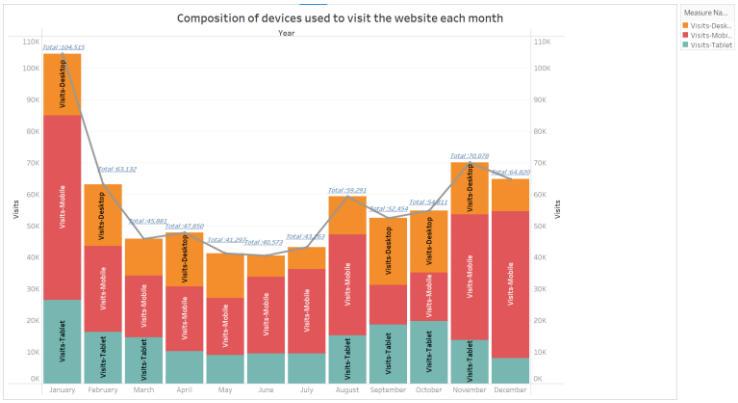


Figure 2—2 Segmented Bar Chart of Device Type Usage Composition Per Month

With the development of the mobile technology, a desktop computer is no longer the only way for people to access the internet. However, for online retailers, this means they must optimise their official site base on the different user interface, so that potential customer can have more fluent shopping experience. The problem with this is that cost of optimization is high while the return sometimes may not compensate the expenditure. Therefore, the company must decide which device optimization should be use more resources than the rest. In this case, the website visits contributed by mobile phones is consistently higher than the rest of the devices. More importantly, the proportion of visits done on desktop and tablets are shrinking by the end of the time period. So as a result, tracking the trend of the major device usage is critical as it will affect the internal resource distribution.

2.3 KPI: Bounce Rate with respect to gender and age band.

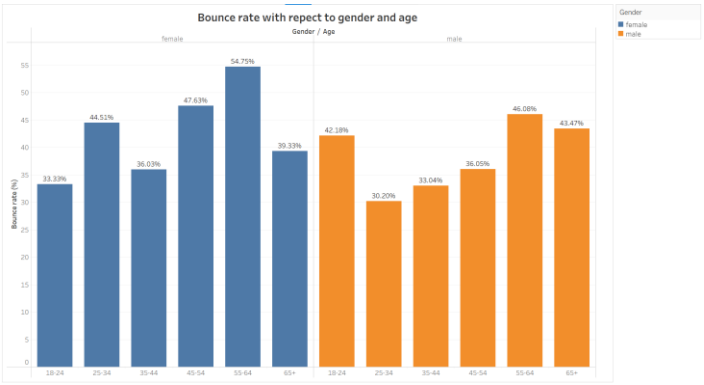


Figure 2—3 Bar Chart of Bounce Rate With Respect to Gender and Age Segmentation

As an online footwear retailor, our content must correspond to the desires of the potential audiences otherwise they may “bounce off” the website after the initial visit. However, people are attracted to different kinds of content with respect to their age and gender. A higher bounce rate for a particular age and gender band means the content of website is not up to their taste, therefore resulting in a less likelihood for purchase to happen and loss of a potential unexploited market growth point.

Part 3 Textual Sentiment Analysis.

To understand want customers desired from the opponent seller, dozens of comments was collected from Instagram which is common virtual place for younger generations to discuss about shoes. 4 competitors were identified (End, Clarks, Schuch, Size?) based on the availability of the comments in the posts and the similarities in operations and size. Over 800 comments in HTML format (569 of which is useable because some of them is emoji) were collected and then refined in Alteryx so that only the text comment were extracted. Lastly, it analysis the wording with each comment and use complex algorithm to determine its sentiment before rearranging the most frequent ones onto the work cloud.

Record	sentiment_category	Count
1	negative	31
2	neutral	488
3	positive	65

Table 3—1 Distribution of Comments at Different Sentiment Levels

The table above displayed overall count of comment based on different sentiment category. It is noticeable that shoe lovers is mostly an up-lifting community because negative comments only occupies a fraction of the cohort. This is also reflected in the world cloud shown below where positive phrase like “love” and “yes” appears most frequently.

Figure 3—1 Word-Cloud Entire Comment

Furthermore, Alteryx can divide the entire data base into 3 different topic group based on the similarities. The first topic word-cloud is shown below, the intended theme should be customer’s expectation of shoes. In which, the most noticeable property of the footwear is the colour blue, followed by yellow and orange. Other than this, price and style are important recurring element in the comments.

Figure 3—2 Word Cloud Topic 2

Figure 3—3 World Cloud Topic 1

Figure 3—4 Word Cloud Topic 3

The second word-cloud is heavily themed in geographical location. The most prominent one is Birmingham, this indicate the demand from the city is strong and need to be filled. What's more, outside of England, the capital city of Wales and Northern Ireland also indicate a strong presence. Lastly, the final topic group seems to be based on the theme of foreign language.

Part 4 Recommendation and Application of Big Data

4.1 Summary off previous findings.

This data analytic projects utilized 2 softwares, Alteryx and Tableau. The former is heavily implemented as a data processing media a which aid the operation by selecting, filtering and correlating data set. In contrast, Tableau is a data visualization tool, which organize data into presentable forms. Overall, linear regression models are more representable than that of logistical. A linear regain model with profit as target variable is the preferable one. In which, quantity and discount used per purchase are decently strong predictors. Unfortunately, they are in a negative correlation with profit which means the more item bought per purchase the less profit the order generate. Such situation is not healthy, and the recommendation is to reduce the discount magnitude.

Another linear model is constructed with sum_quantity purchased overtime (per customer) and it is strongly correlated with the sum of “networking” variables per customer. This indicated that the amount of item purchased by a customer overtime increased as the total time and effort he invested in reviewing websites. This hints to the company that it needs to pay more attention to the web user experience to increase the traffic which may improve the purchase over time.

The research discovers most of the keywords used by potential buyers are related to women's winter wear and children's shoes. It indicates the 2 directions in which the company should pay attention to increasing its online presence. The company also need to pay attention to shifting its focus to the mobile phone platforms as it is the only device which is expanding in usage by potential customers. This could be done by paying close attention to the website optimization on those devices.

Lastly, from the social media comment analysis, it is noticeable that potential audience of the competitors often focuses on the style, colour, and price of the shoes. Most comments also have a geographical property which Birmingham was mentioned most frequently which means there is a potential market to be filled. This is especially true considering our sales in the south-west is not competitive.

4.2 Big Data Application.

Wang and Wang (2020) refer to big data as a way to acquire data with more varied aspects in a faster period of time. It emphasises on volume, variety and velocity (Wolf, 2014). For the volume aspect, it can help us generate more accurate models with a large amount of data. Since Data Set B offers most of the data points monthly, with help of big data, the frequency of data collation can be shortened to hours or minutes. Big data also includes many forms data to be collected (variety), whether it is text, number, or Boolean. So it offers a more comprehensive

way to present a story. Lastly, big data heavily relies on internet connectivity, so the information can be updated at an instant. This is beneficial for the company as it can act much faster instead of reviewing afterwards.

5. Reference

1. Wolff, J.G. (2014), “Big data and the SP theory of intelligence”, IEEE Access, Vol. 2, pp. 301-315.

2. Wang, Shouhong & Wang, Hai (2020) Big data for small and medium-sized enterprises (SME): a knowledge management model. Journal of knowledge management. 24 (4), 881–897.

3. Nickerson, C. (2008) Mutual Suppression: Comment on Paulhus et al. (2004). Multivariate behavioural research. 43 (4), 556–563.

6. Appendix

6.1. Linear Model 3 Pearson Correlation Analysis

Pearson Correlation Analysis

Full Correlation Matrix

	Sales....	Quantity	Discount	Profit....	Number.of.Web.Visit.s..Before.Purchase	Time.Spent.on.the.Website.Before.Purchase..Mins.
Sales....	1	0.33387	-0.1142192	0.7961078	0.0312498	-0.018723
Quantity	0.3338684	1	-0.0093098	0.2356133	0.0353284	-0.0072665
Discount	-0.1142192	-0.00931	1	-0.2415511	0.001283	0.0054369
Profit....	0.7961078	0.23561	-0.2415511	1	0.0224842	-0.0474693
Number.of.Web.Visit.s..Before.Purchase	0.0312498	0.03533	0.001283	0.0224842	1	0.0161189
Time.Spent.on.the.Website.Before.Purchase..Mins.	-0.018723	-0.00727	0.0054369	-0.0474693	0.0161189	1
Number.of.Pages.Visited.Before.Purchase	0.006984	0.00832	0.0069795	0.0229885	0.0043578	-0.0075221

	Number.of.Pages.Visited.Before.Purchase
Sales....	0.006984
Quantity	0.0083235
Discount	0.0069795
Profit....	0.0229885
Number.of.Web.Visit.s..Before.Purchase	0.0043578
Time.Spent.on.the.Website.Before.Purchase..Mins.	-0.0075221
Number.of.Pages.Visited.Before.Purchase	1

6.2. Linear Model 3 Pearson Correlation Analysis

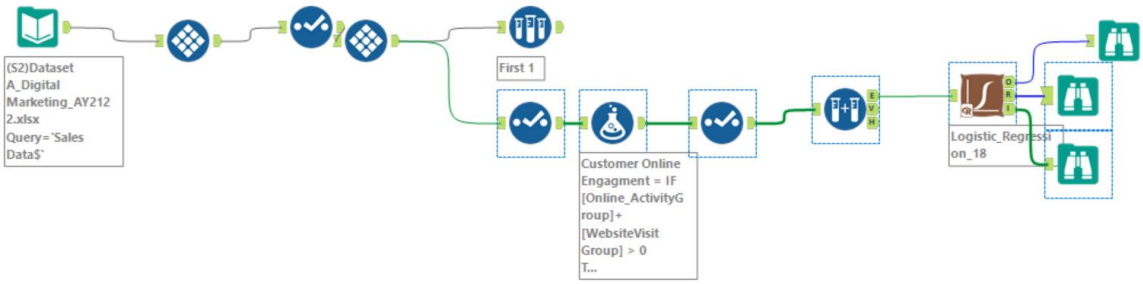
Matrix of Corresponding p-values

	Sales....	Quantity	Discount	Profit....	Number.of.Web.Visit.s..Before.Purchase	Time.Spent.on.the.W
Sales....		0.0000e+00	3.7054e-06	0.0000e+00	2.0689e-01	
Quantity	0.0000e+00		7.0697e-01	0.0000e+00	1.5358e-01	
Discount	3.7054e-06	7.0697e-01		0.0000e+00	9.5868e-01	
Profit....	0.0000e+00	0.0000e+00	0.0000e+00		3.6387e-01	
Number.of.Web.Visit.s..Before.Purchase	2.0689e-01	1.5358e-01	9.5868e-01	3.6387e-01		
Time.Spent.on.the.Website.Before.Purchase..Mins.	4.4960e-01	7.6920e-01	8.2623e-01	5.5128e-02	5.1510e-01	
Number.of.Pages.Visited.Before.Purchase	7.7793e-01	7.3679e-01	7.7807e-01	3.5321e-01	8.6032e-01	

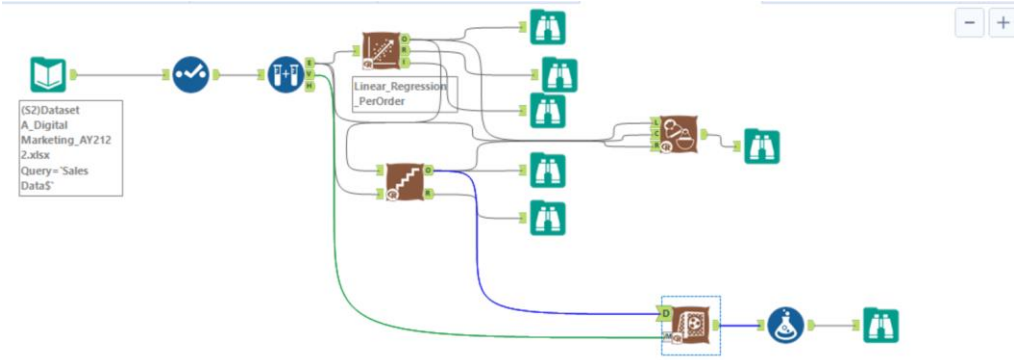
	Number.of.Pages.Visited.Before.Purchase
Sales....	7.7793e-01
Quantity	7.3679e-01
Discount	7.7807e-01
Profit....	3.5321e-01
Number.of.Web.Visit.s..Before.Purchase	8.6032e-01
Time.Spent.on.the.Website.Before.Purchase..Mins.	7.6133e-01
Number.of.Pages.Visited.Before.Purchase	

6.3 Classification Modeling Results and Procedure

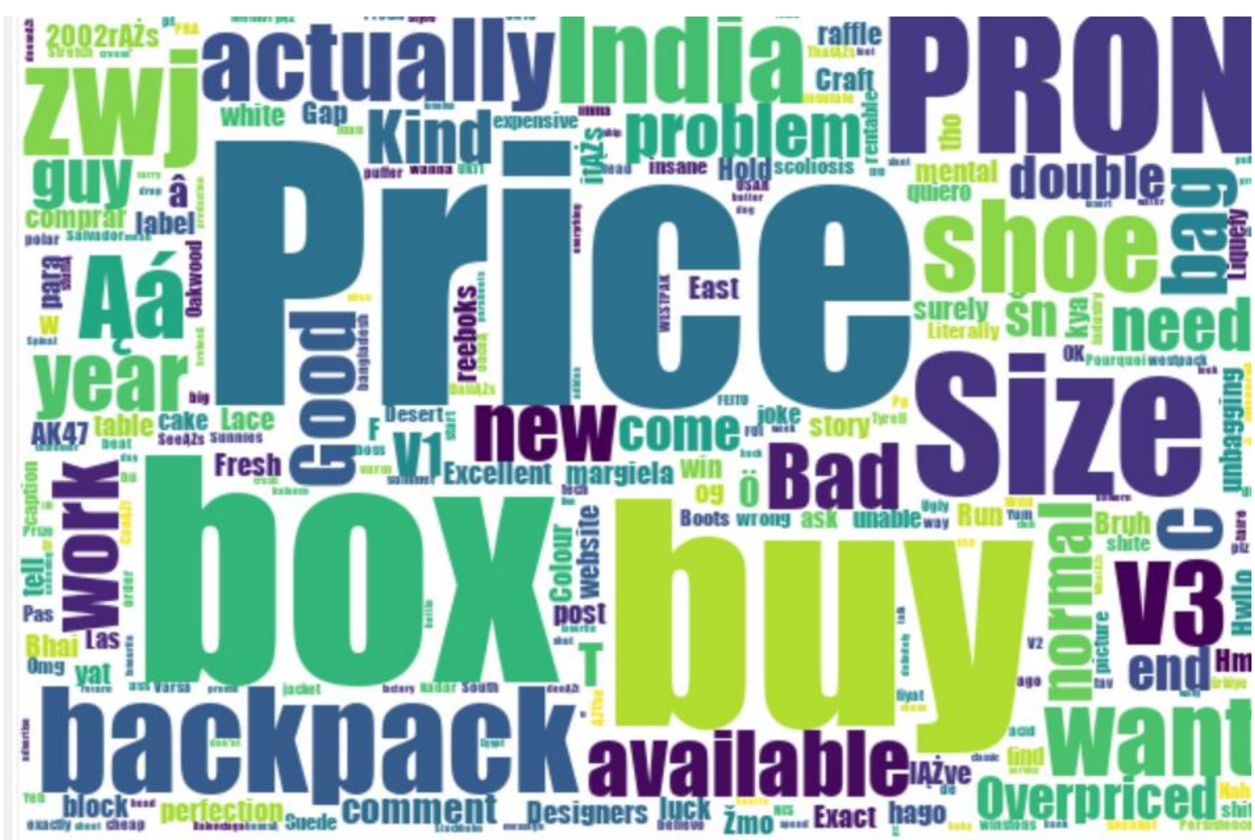
Record	Report
1	**Report for Logistic Regression Model Logistic_Regression_18**
2	*Basic Summary*
3	Call: glm(formula = Responded.to.Email.Campign. ~ Number.of.Web.Visit.s..Before.Purchase + Time.Spent.on.the.Website.Before.Purchase..Mins. + Number.of.Pages.Visited.Before.Purchase + Loyalty.Card.Holder.. + Downloaded.Discount.Voucher.for.Future.Purchase. + Customer.Online.Engagment, family = binomial("logit"), data = the.data)
4	Deviance Residuals:
5	Min 1Q Median 3Q Max -1.39 -1.18 1.00 1.16 1.34
6	Coefficients:
7	(Intercept) 0.074222 0.346428 0.2142 0.83035 Number.of.Web.Visit.s..Before.Purchase 0.072443 0.045776 1.5825 0.11353 Time.Spent.on.the.Website.Before.Purchase..Mins. -0.001027 0.005424 -0.1893 0.84986 Number.of.Pages.Visited.Before.Purchase -0.020668 0.014467 -1.4286 0.15311 Loyalty.Card.Holder..Yes -0.215194 0.119072 -1.8072 0.07072 . Downloaded.Discount.Voucher.for.Future.Purchase.Yes -0.078908 0.118938 -0.6634 0.50705 Customer.Online.EngagmentLow Online Engagment Custome 0.114268 0.162358 0.7038 0.48155 Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for binomial taken to be 1)
8	Null deviance: 1584 on 1142 degrees of freedom Residual deviance: 1575.9 on 1136 degrees of freedom McFadden R-Squared: 0.005098, Akaike Information Criterion 1590

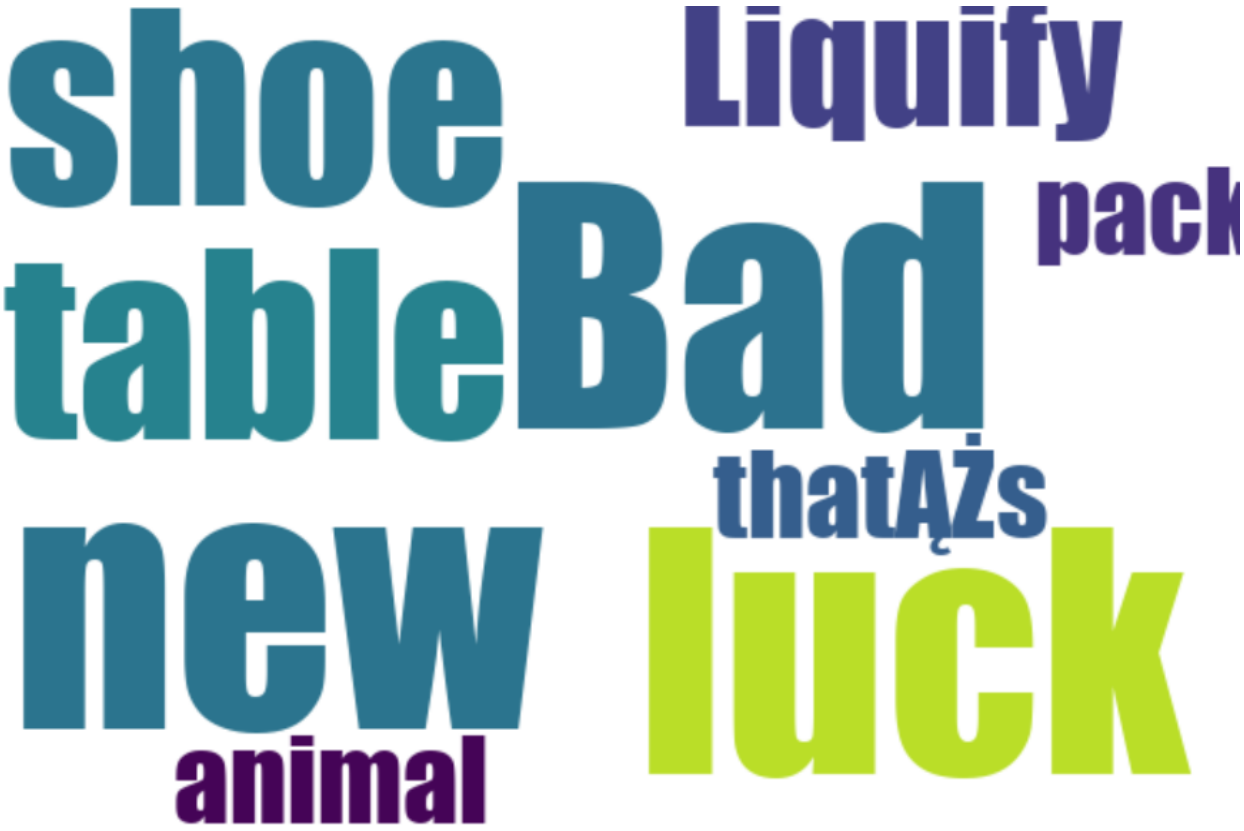


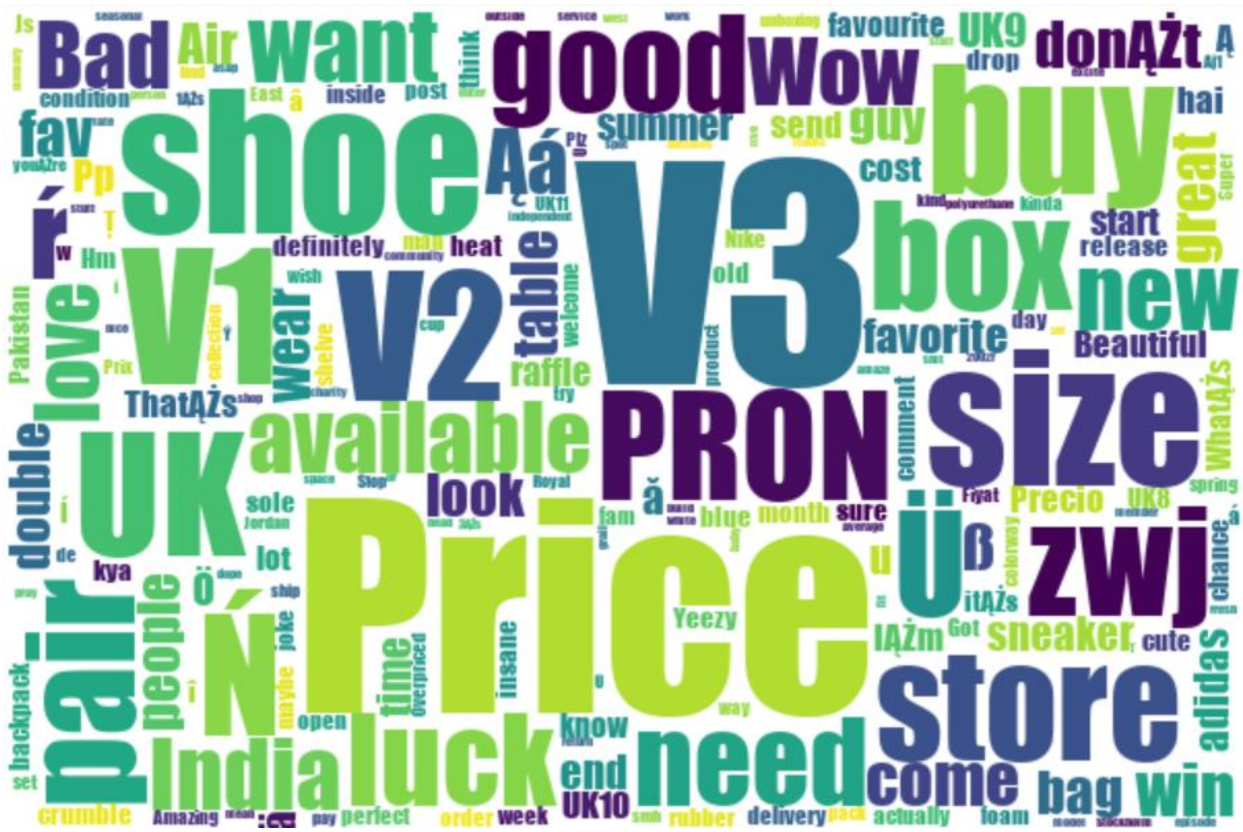
6.4 Linear Regression Model #1 Procedure.




shoe
UK
pair
V3
size
store
buy
Price
good
PRON
available
need
look
love
great
Adidas
time
luck
sneaker
Beautiful
DonĄŻt
win
India
people
wear
want
Wow
UK9
favorite
come
Air
bag
Yeezy
raffle
chance
Pakistan
release
delivery
crumble
send
know
sole
blue
summer month
collection
condition
zwj


Price
PRON
India
actually
zwj
Kind
problem
shoe
double
bag
need
Size
box
buy
backpack
available
want
Overpriced
new
come
Good
Bad
work
year
guy
V3
V1
normal
comment
perfection
Designers
block
Excellent
margiela
reebkos
label
table
cake
Lace
Fresh
AK47
unbagging
Hwllo
surely
Literally
East


shoe
Liquify
pack
table
Bad
thatĄŻs
new
luck
animal


Price
V3
V2
V1
shoe
good
buy
size
PRON
store
UK
pair
luck
India
need
available
box
want
Bad
Wow
zwj
new
love
come
wear
table
look
people
great
favorite
Air
fav
summer
guy
send
cost
start
release
Beautiful
raffle
time
double
backpack
crumble
Amazing
perfect
order
week
UK10
rubber
delivery
actually
foam
bag
win
adidas
sneaker
cute
Yeezy
insane
open
know
end
joke
lot
sole
kya
month
sure
blue
Precio
UK8
comment
chance
WhatĄŻs
spring
definitely
heat
inside
post
drop
hai
condition
East
favourite
UK9
donĄŻt
Pakistan
Hm
Pp
welcome
think