IST 707 – Final Project Report
E-Commerce Analysis
Abhiram Gopal & YuanyuanXue

Contents

Module 1: Introduction and Problem Statement 2

Module 2: Data Assimilation and Understanding 2

Module 3: Cleaning and Manipulation 3

Module 4: Exploratory Data Analysis 3

Module 5: Model Creation 6

Module 6: NLP Word Cloud Analysis 7

Module 7: Business/ Statistical Inference 8

Appendix

Module 1 Introduction and Problem Statement

Nowadays, E-commerce is becoming more and more popular. Many people prefer shopping online to shopping onsite. In this project we are trying to measure the performance of products in E-commerce site "WISH" during the month of summer. This analysis can be useful for any merchant who wants to list their product on the "WISH" ecommerce site specially during the month of summer.

Major data Questions:

Our primary objective is to answer the following question through this project:

- Try to predict the customer behavior using the feature coefficients & their significance.
- Customer buying preferences in the WISH ecommerce site
- Keywords & Tags in items description which the merchant can use

Module 2: Data Assimilation and Problem Understanding

- This dataset contains product listings as well as products ratings and sales performance of product listings in the WISH platform in August-2020 with search term "Summer Sales"
- The dataset roughly contains around 1500 rows & 43 columns.(Illustration 1.1)
- The dataset contains many unwanted columns, incorrect entries such as highlighted(product color, size, variation, country..etc)(Fig 2.1)
- Target class: The Y variables defines the quantity of units sold for a product – The 'units_sold' Variable(Fig 2.2)

We have conceived this problem as a more inferential type of analysis wherein our focus is to find the factors responsible for higher quantities of units_sold. In order to do this, we shall be planning on building a regression model. We shall also discretize the units_sold column and use a decision tree. We shall then be using the feature importance scores of each model & check if they both compliment each other. Finally, with the help of these factors we shall provide a business solution targeted towards new merchants who want to list their product in the WISH website specially during the summer month.

```
> str(df_new4$product_color)
 Factor w/ 102 levels "","applegreen",..: 91 38 46 9 102 61 91 17 9 8 ...
```

Fig 2.1 : The product color containing 102 levels

```
> summary(df_new6$units_sold)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1     100    1000    4422    5000  100000
```

Fig 2.2 : The Dependent Variable is a numeric type

Module 3: Cleaning and Manipulation

1. **Product_color**
   - The product color contains 102 levels. The colors are highly specific wherein a simple color of black has 4 variations - “black”,”offblack”,”coolblack”,”Black”. SImilarly for all others colors like blue, red, green, multicolor..etc
   - Essentially, I have converted these specific color variations into broader one.
   - Now I have 15 color variations containing the most broadest colors of spectrum.

2. **Product Variation size**
   - Similar to the product_color, even this column had like 107 levels but most of its entries were redundant & incorrect entries.(Such as ‘XxL’ instead of ‘XXL’)
   - Converted those incorrect entries into proper ones by string manipulations.(converting all to lower case & also defined sizes from XXXXXS to XXXXXL.

3. **Duplicate Rows & ID columns**
   - The dataset contained a lot of duplicate rows & also ID columns which totally does not affect the performance of our model nor does it help in understanding our dataset.
   - SO we shall be removing those ID columns & also duplicate rows in the dataset.

4. **Shipping_option_name**
   - Similar to the previous ones, we compacted the shipping column(of 15 lvls) into 2 factors of “Standard Shipping or Express Shipping”

5. **NULL Value Interpolation:**
   - We found that there are a few missing values in our data. We can simply delete rows with missing values, but usually we would want to take advantage of as many data points as possible. Replacing missing values with zeros would not be a good idea. So, we have used the "replace_na" of tidyr to fill every NaN (not a number) entry with the mean of the column

Module 4: Exploratory Data Analysis

1. There seems to be an interesting relation where the White & Black colored products seems to be selling the most.(Illustration 4.1)
2. There’s also a relation where product sizes of small & medium sells the most compared to other product sizes (Illustration 4.2)
3. The products having high sales have their retail prices below 1000$(Illustration 4.3)
4. As the merchant begins to sell more & more products, his rating automatically increases & averages between 3 & 5.(Illustration 4.4)

Variable Reduction:

- There are a whopping 43 variables in our dataset. There is no way we would be able to use all of these variables for our predictive models.
  - Dropping Variables with more than 50% values missing
    1. Removed all the variables which have more than 50 % missing values.
    2. There was 1 column with more than 50 % missing values

- Dropping variables with very less variance & ID type columns
    1. Used a 10 % threshold for variance.
    2. Dropped variables which are below 10 %
    3. There were a total of 7 variables below 10 %

	Number of Columns	**Number of Rows**
Before Transformation	43	1573
After Transformation	18	1539

Module 5: Model Creation

Linear Regression

Before we begin with the linear regression, we created dummy columns for all factor types so as to essentially capture the relationships. We also used the correlation matrix to remove those numerical columns having high correlation(with a threshold of 90%). In this project we shall be using the "lm" function to plot the linear model.(Illustration 5.1)

From our illustration :-

1. Retail price, rating count,product variation size & merchant rating count are the most significant in influencing the quantity of units sold
2. Product color, shipping name and origin countries may have little influence in the unit sold.
3. As also explored in the EDA section, the sizes XXS & XXL is significant for predicting the units_sold.
4. Most other columns have either very negligible coefficient scores or very high P-value which means that those columns are not contributing much to our model..

Now that we have our results of our Linear Regression, We shall be trying out decision tree & to check whether the results from decision tree & linear model support each other or not.

Decision Tree

Before we begin with our decision tree, we first discretized the units_sold Y column. We split the units sold into 3 categories based upon the distribution of the column. Decision Tree is basically a "IF-ELSE" algorithm that can be used to find the top nodes or the top "IF" conditions which are essentially the features that we are looking for. For our Decision Tree we shall be using the CARET package.(Illustration 5.2, Illustration 5.3)

From our illustration :-

1. A rating count of more than 1500 is required if you want to have sales greater than 5000.
2. It can be observed that Both 5 star rating & one star ratings play a really important role in determining the number of units sold.
3. Merchant Rating & product variation inventory has a very major role in determining the quantity of units_sold.

Module 6: NLP Word Cloud Analysis

For this particular module however, we are using Python for the operation. Our objective here is to analyze those strings of "Tags","title" columns of products which have a large number of quantities sold. So, first we are filtering the data frame having only these columns. Then we shall tokenize the entire column & use the word cloud feature to plot the word cloud.(Illustration 6.1)

From the Illustration:

1. Top selling tags usually include women,fashion. This can also lead to a conclusion that most of the customers are usually women shopping during the summer month.
2. "Plus size" , "Sleeveless" are also really common tags and they too attract a lot of sales & attention from women customers.

Module 7: Business/Statistical Inference:

- White, black & blue colors of products seems to be selling well(based on visual Analysis)
- Any merchant should have good amounts of Small & Medium size Apparel
- Rating count & merchant rating count is very significant in influencing the units sold.(based upon the coefficients in Linear Regression)
- You should be sure that you ensure good customer satisfaction because the product rating count & merchant rating count plays a significant role in determining your sales.(Also verified in Decision Tree as well)

APPENDIX

1. **Dataset summary**

Illustration 1: Dataset Summary & columns with incorrect entries highlighted

2. EDA ILLUSTRATIONS:

Illustration 4.1: Colours vs Units Sold

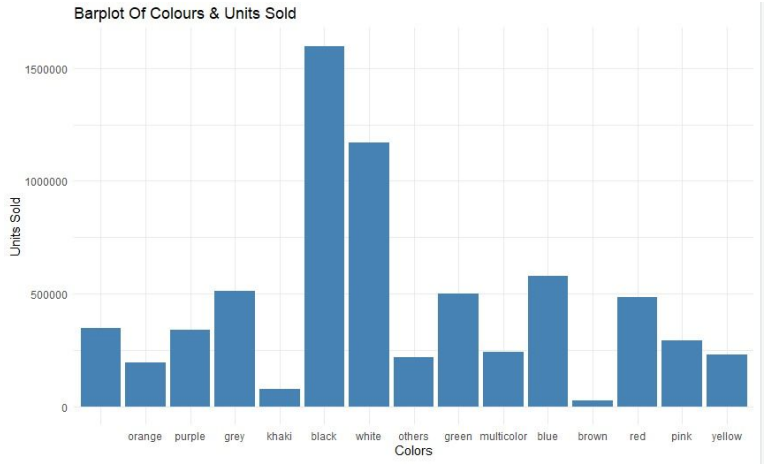

Illustration 4.2 : Product Size vs Units Sold

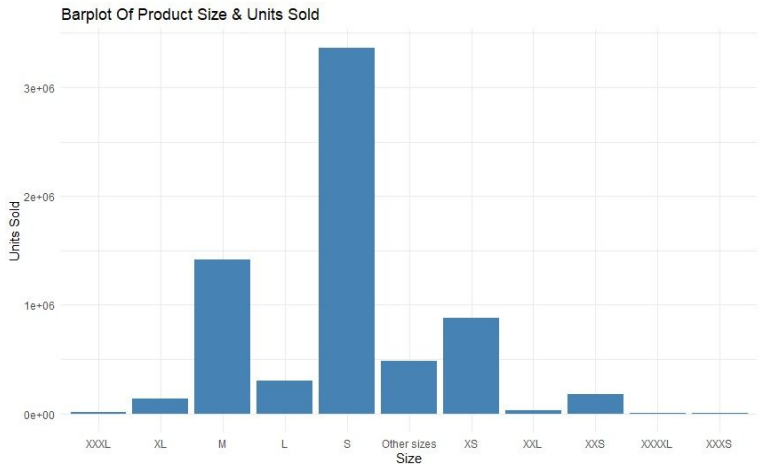

Illustration 4.3 : Retail Price vs Units Sold

Illustration 4.4 : Retail Price vs Units Sold

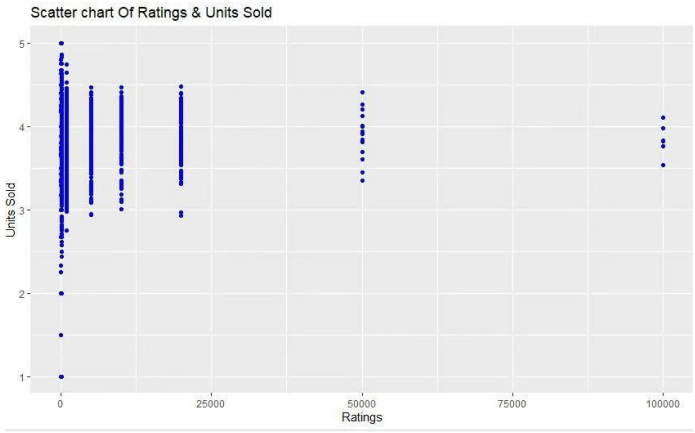

Illustration 5.1: Linear regression

```
Call:
lm(formula = units_sold ~ ., data = df_new6)

Residuals:
    Min      1Q  Median      3Q     Max
-28719  -1119    -287     616   49405

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.016e+03  2.739e+03   1.831 0.067259 .
price                            1.565e+01  5.769e+01   0.271 0.786229
retail_price                    -1.234e+01  3.642e+00  -3.388 0.000723 ***
uses_ad_boosts1                  4.925e+02  2.136e+02   2.305 0.021285 *
rating                          -1.379e+01  2.211e+02  -0.062 0.950293
rating_count                     6.184e+00  9.077e-01   6.813 1.38e-11 ***
rating_five_count               -3.565e+00  1.367e+00  -2.608 0.009197 **
rating_one_count                -3.210e+00  2.779e+00  -1.155 0.248126
product_colorgreen              -5.655e+01  7.601e+02  -0.074 0.940701
product_colorothers             -1.982e+02  8.767e+02  -0.226 0.821144
product_colorblack              -9.903e+02  7.089e+02  -1.397 0.162645
product_colormulticolor         -8.328e+02  8.451e+02  -0.985 0.324560
product_colorblue               -2.697e+02  7.483e+02  -0.360 0.718631
product_colorbrown               4.821e+02  1.708e+03   0.282 0.777730
product_colorred                -4.912e+02  7.573e+02  -0.649 0.516642
product_colorpink               -5.689e+02  7.942e+02  -0.716 0.473891
product_colorgrey               -5.694e+02  8.068e+02  -0.706 0.480489
product_colorkhaki              -4.888e+02  1.308e+03  -0.374 0.708754
product_colorpurple              7.902e+02  8.668e+02   0.912 0.362120
product_coloryellow             -1.115e+03  7.937e+02  -1.405 0.160122
product_colorwhite              -3.138e+02  7.169e+02  -0.438 0.661658
product_colororange              1.422e+03  1.034e+03   1.375 0.169480
product_variation_size_idM      -8.230e+02  6.516e+02  -1.263 0.206753
product_variation_size_idother sizes -2.338e+03  6.620e+02  -3.531 0.000426 ***
product_variation_size_idS      -1.657e+03  6.072e+02  -2.729 0.006427 **
product_variation_size_idXL     -1.816e+03  1.158e+03  -1.569 0.116932
product_variation_size_idXS     -1.924e+03  6.398e+02  -3.007 0.002679 **
product_variation_size_idXXL    -4.095e+03  1.219e+03  -3.358 0.000804 ***
product_variation_size_idXXS    -2.659e+03  7.498e+02  -3.546 0.000403 ***
product_variation_size_idXXXL    1.880e+03  4.109e+03   0.458 0.647276
product_variation_size_idXXXS   -2.851e+03  1.781e+03  -1.601 0.109576
product_variation_size_idXXXXL  -2.838e+03  2.932e+03  -0.968 0.333251
product_variation_inventory      6.817e+00  5.752e+00   1.185 0.236112
shipping_option_name1            1.091e+03  4.816e+03   0.227 0.820823
shipping_option_name2           -3.214e+02  9.770e+02  -0.329 0.742207
shipping_option_price           -3.332e+02  2.240e+02  -1.488 0.136983
shipping_is_express1            -5.446e+02  4.367e+03  -0.125 0.900781
countries_shipped_to            -9.129e+00  5.386e+00  -1.695 0.090326 .
origin_countryOther Countries   -9.175e+02  8.654e+02  -1.060 0.289212
origin_countryUS                 4.258e+01  7.475e+02   0.057 0.954584
merchant_rating_count            5.490e-03  1.410e-03   3.893 0.000103 ***
merchant_rating                 -2.589e+02  5.566e+02  -0.465 0.641842
---
```

Illustration 5.2

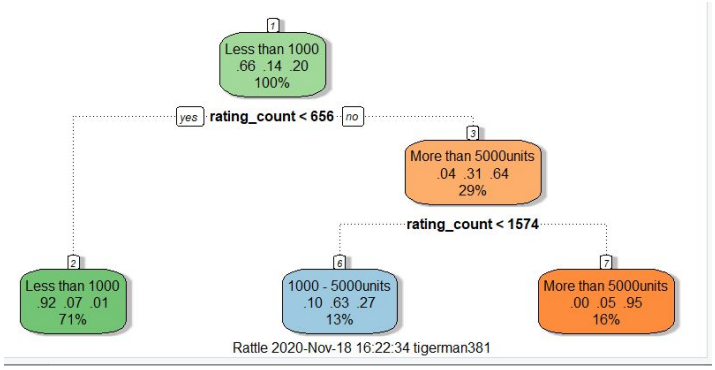

Illustration 5.3

only 20 most important variables shown (out of 41)

	Overall
rating_count	386.712
rating_five_count	367.272
rating_one_count	315.989
merchant_rating_count	62.312
product_variation_inventory	16.939
rating	4.378
product_colorothers	0.000
product_colorgreen	0.000
origin_countryUS	0.000
product_colorbrown	0.000
product_colorwhite	0.000
product_colorpink	0.000

Illustration 6.1