Final Report: Statistical Modeling and Analysis Results for the Mercari Shopping App Price Suggestion Project

Submitted to:

Mentor,

Data Science

Edwisor

Report Prepared By:

Sugand. A

March 23, 2018

Table of Content

Executive Summary

This report summarizes the statistical modeling and analysis results associated with the Price Suggestion of products for the Online Shopping App MERCARI.The purpose of this report is to document both the implemented design and all corresponding data modeling and inference techniques used during the subsequent statistical analyses. Programming language used was R through a kernel in Jupter notebook.The evaluation metric used is Root Mean Squared Logarithmic Error as it can penalize the under estimates more than over estimates.

1.0 Problem Statement

It can be hard to know how much something's really worth. Small details can mean big differences in pricing. For example, one of these sweaters cost $335 and the other cost $9.99. Can you guess which one's which?

Sweater A:

"Vince Long-Sleeve Turtleneck Pullover Sweater, Black, Women's, size L, great condition."

Sweater B:

"St. John's Bay Long-Sleeve Turtleneck Pullover Sweater, size L, great condition"

Product pricing gets even harder at scale, considering just how many products are sold online. Clothing has strong seasonal pricing trends and is heavily influenced by brand names, while electronics have fluctuating prices based on product specs.

Mercari, Japan's biggest community-powered shopping app, knows this problem deeply. They'd like to offer pricing suggestions to sellers, but this is tough because their sellers are enabled to put just about anything, or any bundle of things, on Mercari's marketplace.

Provided with user-inputted text descriptions of their products, including details like product category name, brand name, and item condition, an algorithm is to be built that automatically suggests the right product prices.

2.0 Data used

As provided, the data is already pre split into train, test and sample submission. The train data contain's 8 variables and 1482535 observations. The test data contain's 7 variables and 693359 observations which excludes price i,e out target variable for which we'll be making the predictions for test data.

3.0 Variables used

1. **id -** the id of the listing
2. **name -** the title of the listing
3. **item_condition_id -** the condition of the items provided by the seller
4. **category_name -** category of the listing
5. **brand_name –** brand name of items
6. **price –** Price in USD(Our target variable)
7. **shipping -** 1 if shipping fee is paid by seller and 0 by buyer
8. **item_description -** the full description of the item**.**

4.0 Data Preprocessing:

Our train and test datasets were loaded and checked for the following,

- **Structure**
- **Summary**
- **Missing values**

After getting the data altogether it has dimension of 8 variables and 1482535 observations. Out of 8 variable's 4 were character(name, brand_name, category_name , item_description)3 were integers(id, item_condition_id and shipping) of which (item_condition_id and shipping)should be treated as categorical for our analysis. Our target "Price " was numerical.

There were found to be no missing values.

5.0 Exploratory Data Analysis

5.1 Analysis of target variable

- **Price**

Checking the distribution of the variable,

Histogram was plotted to check the distribution of the variable.

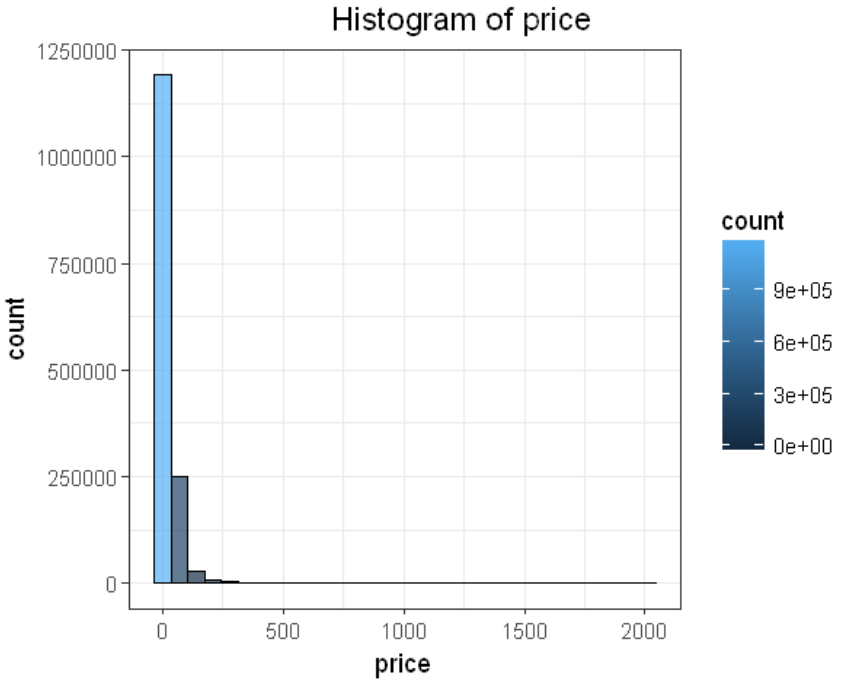


Fig 5.1

Price was very left-skewed. The minimal price is 0, while the highest is 2009. Price was transformed using log(x+1) to get rid of skewness (the +1 is there to avoid taking log(0)).

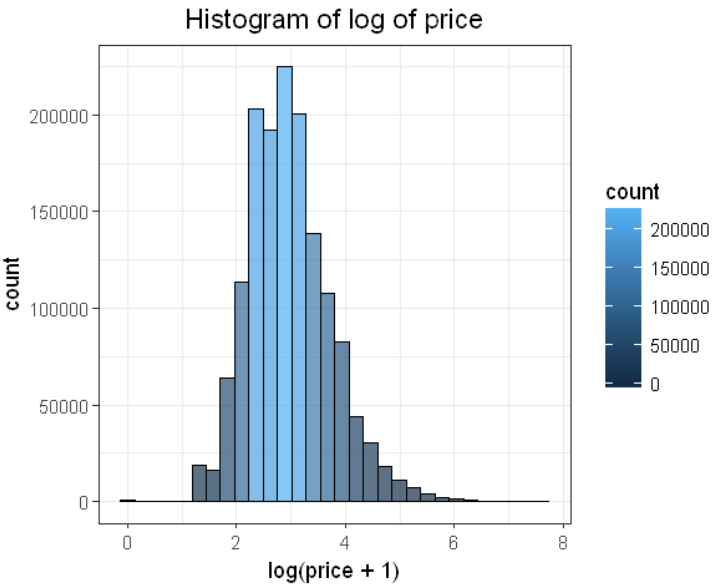


Fig 5.2

Checking for the number of items that were priced '0' there were found to be 874 of them.

Looking into the above, it's quite obvious that either these values are wrong or items are given away for free. Let's analyze that in the later stage. Now let's move on for our next step.

5.2 Analysis of categorical variables

- **item condition id**

The range of item_condition_id was found to be between 1 and 5 of which,

```
   1      2      3      4      5
640549 375479 432161  31962   2384
```

Items with condition_id 1 were 640549 items , 2 were 375479, 3 were 432161, 4 w ere 31962 and 5 were 2384.

Visualization of the same was done using bar plot.

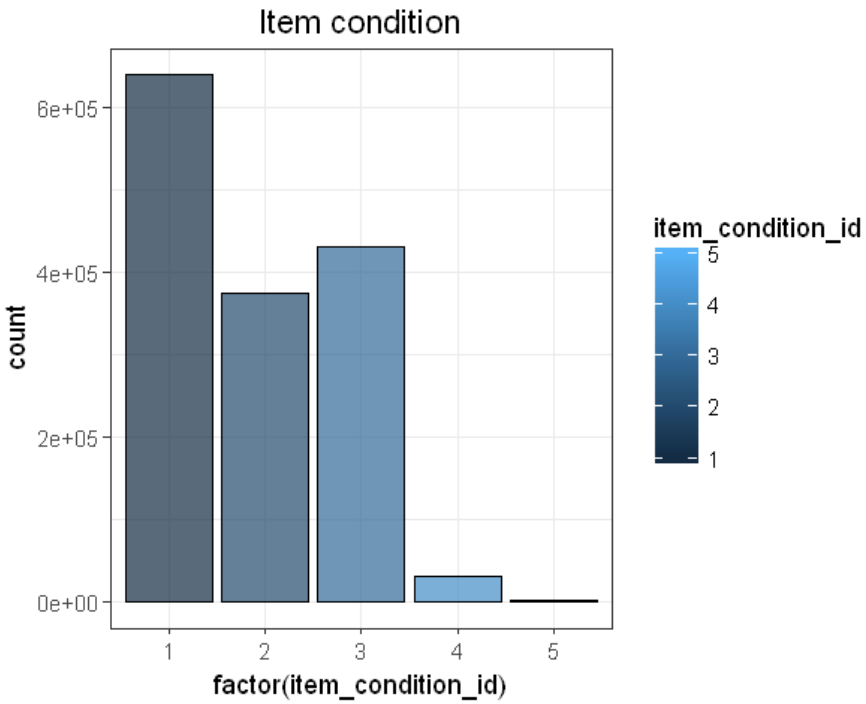


Fig 5.3

Looking at the median prices for item conditions:

item_condition_id	medianprice
1	18
2	17
3	16
4	15
5	19

Very interesting, we'll catch up this analysis later.

- **Shipping data**

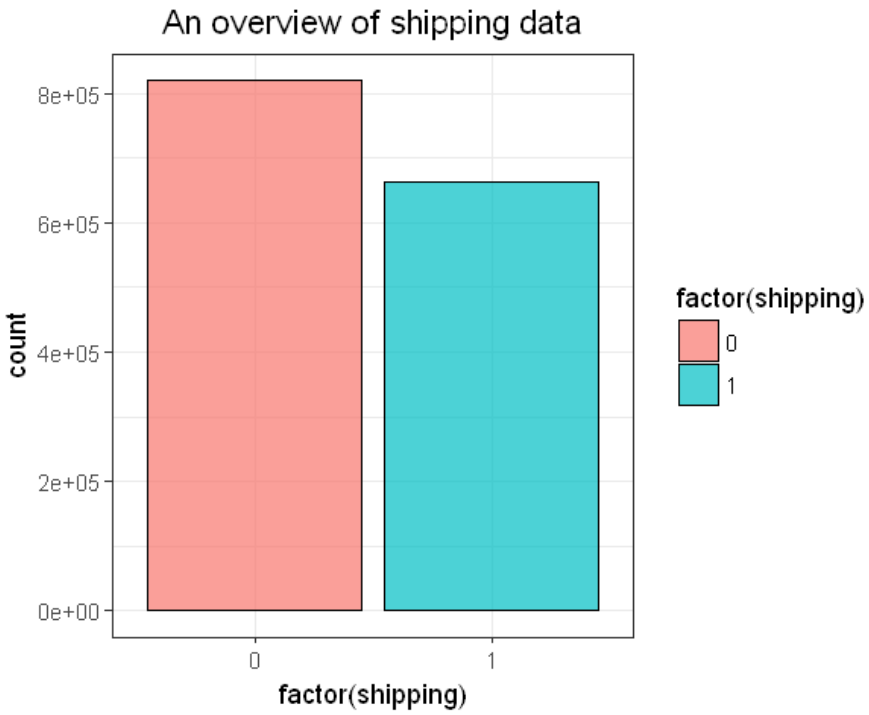


Fig 5.4

1 if shipping fee is paid by seller and 0 by buyer, clearly states that for most of the items ship ping fee is paid by the buyer itself.

5.3 Bivariate analysis

Now lets see the relationship between item condition and price, we already saw the relati on ship between the median prices and item_condition_id, here it is once again.

item_condition_id	medianprice
1	18
2	17
3	16
4	15
5	19

Let's plot the same with boxplot.

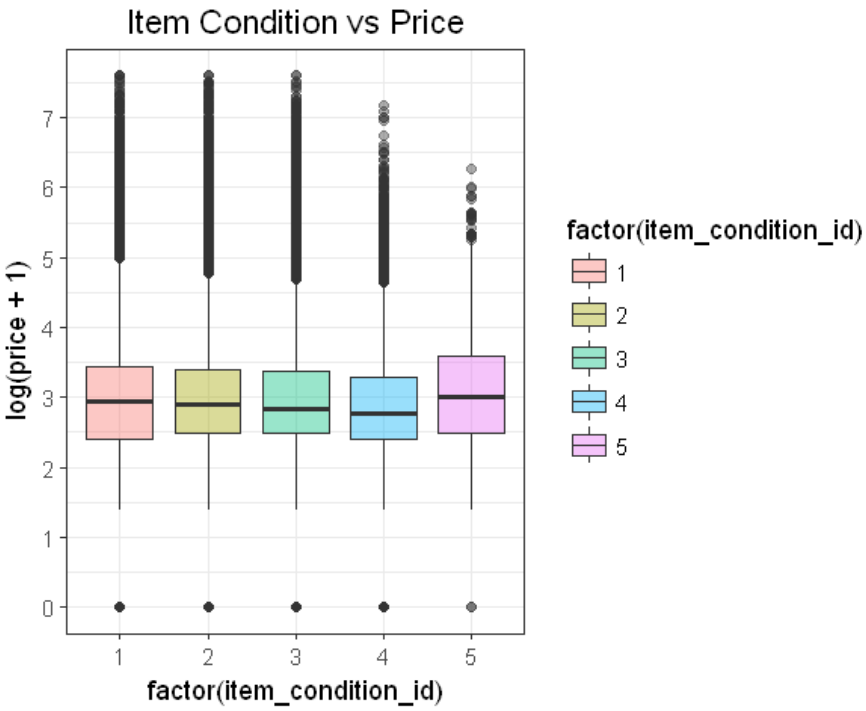


Fig 5.5

Very interesting to see that condition 5 is having the greater price, upon research could find that item condition 5 is the worst, hence we can relate that branded expensive items that are very old are sold, we'll go through this when we analyze our text data.

Now let's see the relationship between price and shipping data.

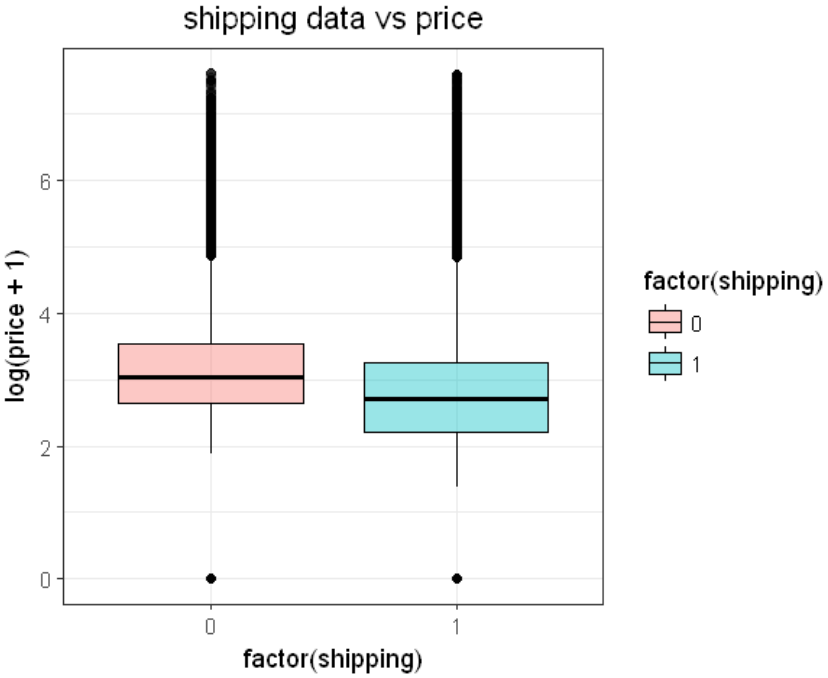


Fig 5.6

We assume that if the shipping fee is paid by seller the price to be highbut again interest ing to see that it for products where the seller isn't paying the shipping fee is high priced this might be due to the product category and branding etc, we'll explore that in a while.

Relationship between price, shipping and item condition_id, Prices were choosen to be g reater than the median i,e 19 and the item_condition_id being the lowest i,e 5 with shipp ibg being 0 i,e paid by the buyer.

name	item_condition_id	category_name	brand_name	price	shipping	item_description
Fossil vintage renewal purse	5	Women/Women's Handbags/Shoulder Bag	Fossil	36	0	No description yet
Gameboy advance sp ags-101 FOR PARTS	5	Other/Other/Other		24	0	For parts . Turns on and hold battery charge . Read the games perfect .
Nintendo for parts	5	Electronics/Video Games & Consoles/Consoles		20	0	Powers on but won't play games.
Kd 6 Aunt Pearl size 8	5	Men/Shoes/Athletic	Nike	36	0	Poor conditon kd 6 Aunt pearl. Really beat. Comes with og box
Gamecube NOT READING LOT	5	Electronics/Video Games & Consoles/Games	Nintendo	22	0	Games do not read. As is condition. Repost of other listing- added 2 more games Rampage and a very very bad condition Pokemon XD. Let me know if you have any questions. For [rm] more I'll include the game cases for Dragonball and Sonic too, just let me know.
Hobo International Lauren Wallet	5	Women/Women's Accessories/Wallets		24	0	Pretty wallet but the back closing clip is broken. Can probably be repaired by a cobbler. Otherwise this is a very pretty wallet. Price is firm and no free shipping!! Thanks

Upon filtering it was found to be that 782 items were in that category but the question was w ere all these branded?

Upon checking for branding it was found that

```
TRUE  FALSE
 624   158
```

So out of 782, 624 are branded and the remaining aren't. So it clearly makes sense that bra nded expensive items that are very old are sold under category 5

5.4 Exploration of Character Class Variable's

- **Category_name**

Checking for the length of unique set of variables in Category_name, it was found that 1288 of them were unique which is a slightly high number.

Checking for top ten of them:

```
                                                  Women/Athletic Apparel/Pants, Tights, Leggings
                                                                                           60177
                                                               Women/Tops & Blouses/T-Shirts
                                                                                           46380
                                                                      Beauty/Makeup/Face
                                                                                           34335
                                                                      Beauty/Makeup/Lips
                                                                                           29910
                                                 Electronics/Video Games & Consoles/Games
                                                                                           26557
                                                                      Beauty/Makeup/Eyes
                                                                                           25215
                                  Electronics/Cell Phones & Accessories/Cases, Covers & Skins
                                                                                           24676
                                                                    Women/Underwear/Bras
                                                                                           21274
                                                                 Women/Tops & Blouses/Blouse
                                                                                           20284
                                                               Women/Tops & Blouses/Tank, Cami
                                                                                           20284
```

Upon splitting these ,There were 6 categories found and top 3 were retained while the other 3 ignored since it contained plenty of missing values adding no meaning to the data.

3 categories were formed and now let's start analyzing them.

For hierarchical categories analysis, treemaps should be the best choice .

1st and 2nd level category analysis.

1st and 2nd Hierarchical Category Levels

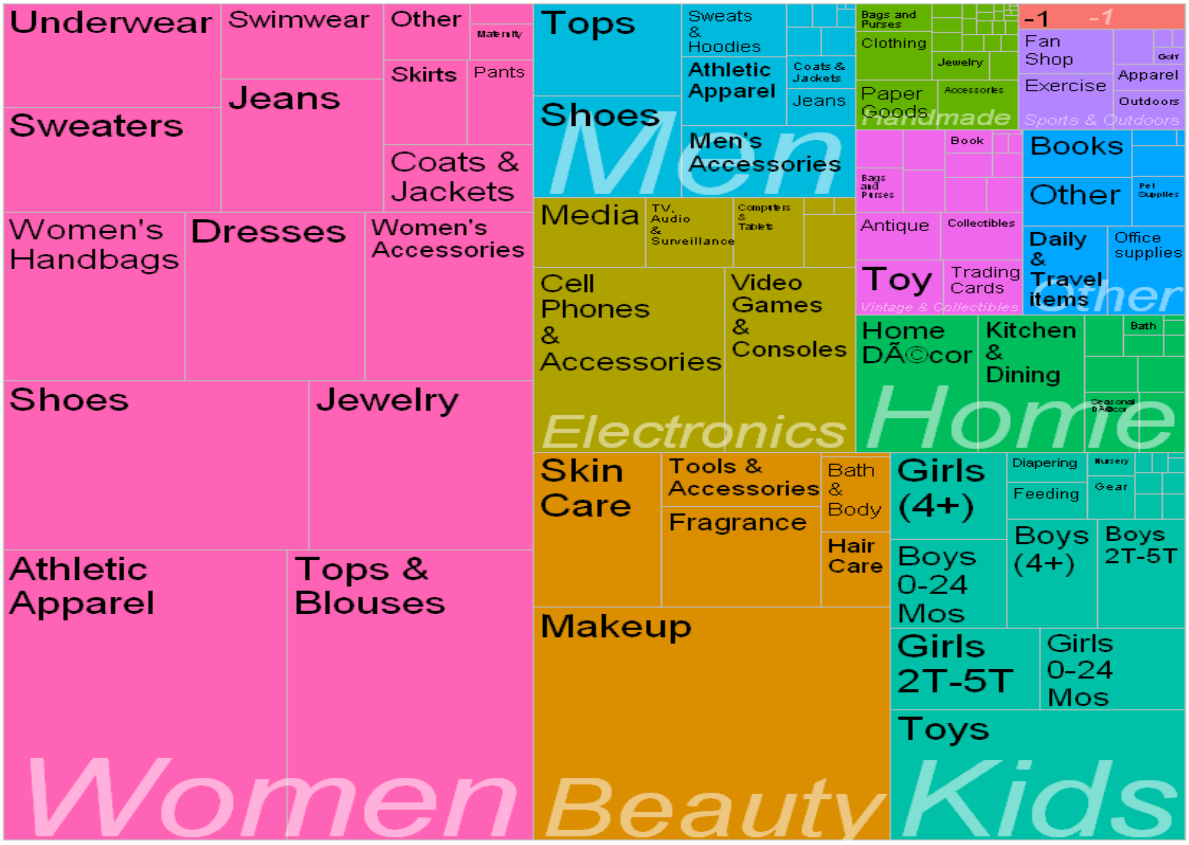


Fig 5.7

It's evident that Women and Men form the most, hence let's start analyzing them too,

2nd and 3rd Hierarchical Category Levels under Women

Fig 5.8

2nd and 3rd Hierarchical Category Levels under Men

Fig 5.9

Analyzing 2nd 3rd level categories:

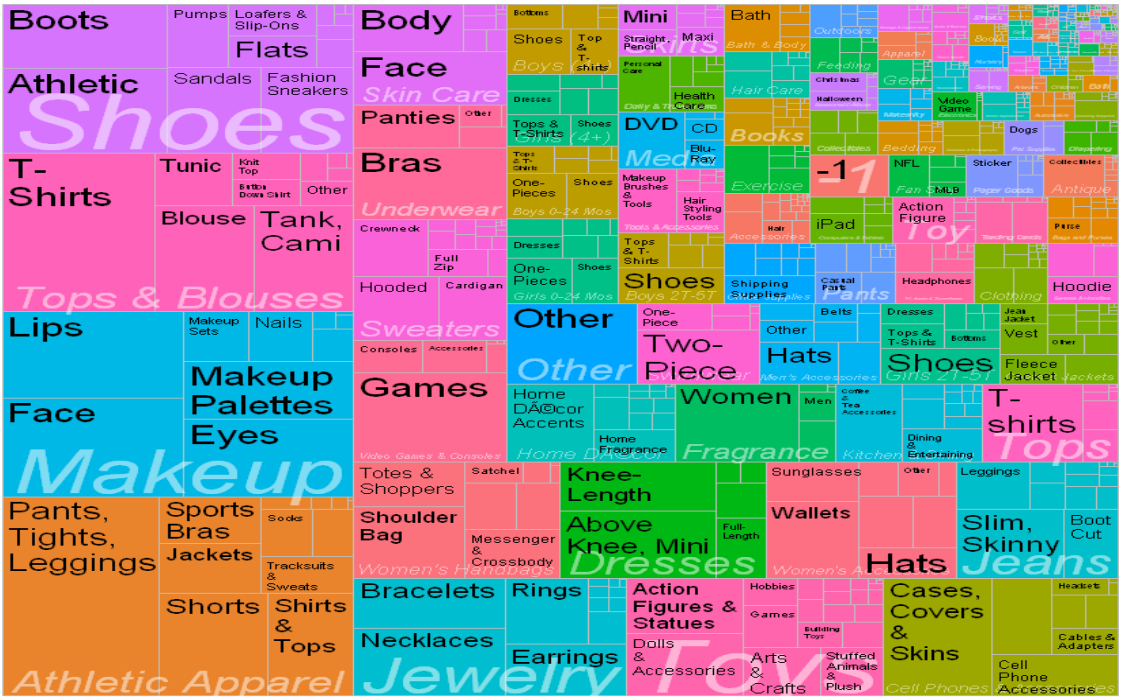


Fig 5.10

We shall a keep a record of the top treemaps for later use.

Now let's analyze the branding,

- **brand analysis**

Now let's look at the 1st level category items with and without brand.

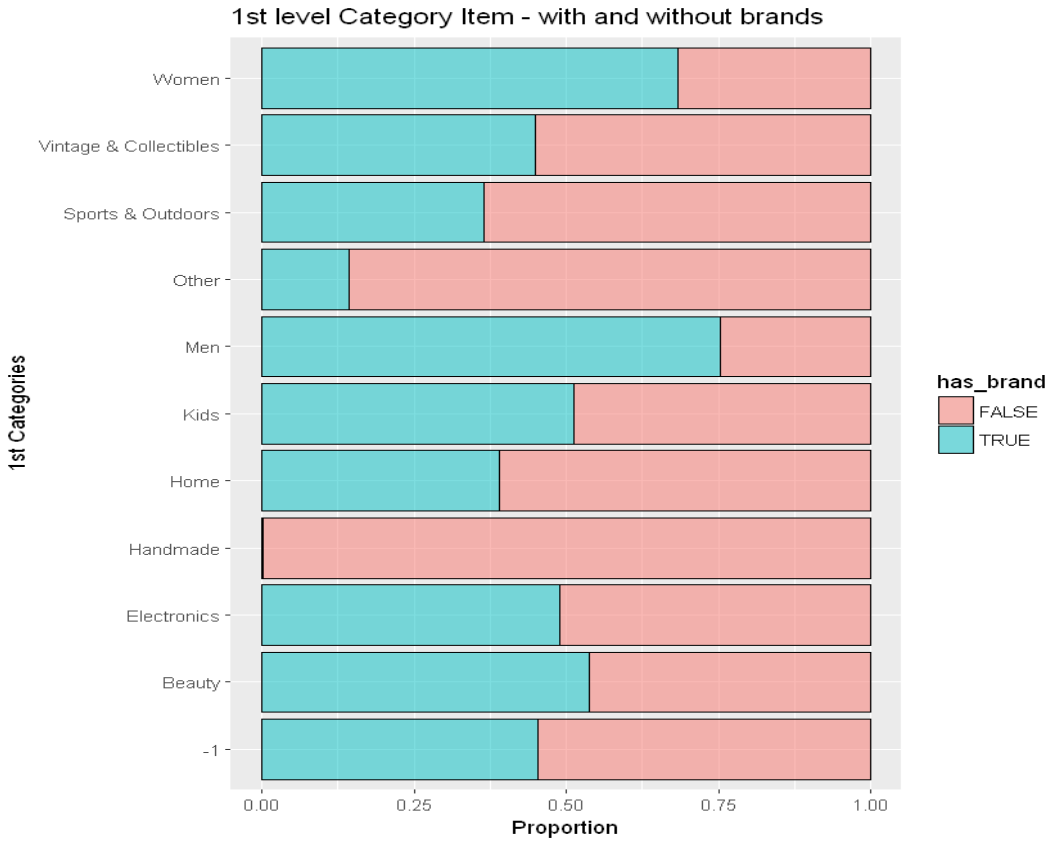


Fig 5.11

We can see that most of the Mens and Womens items have a brand while handmade items do not have one, which clearly makes sense.

Analysis of top 20 most sold brands and their Categories.

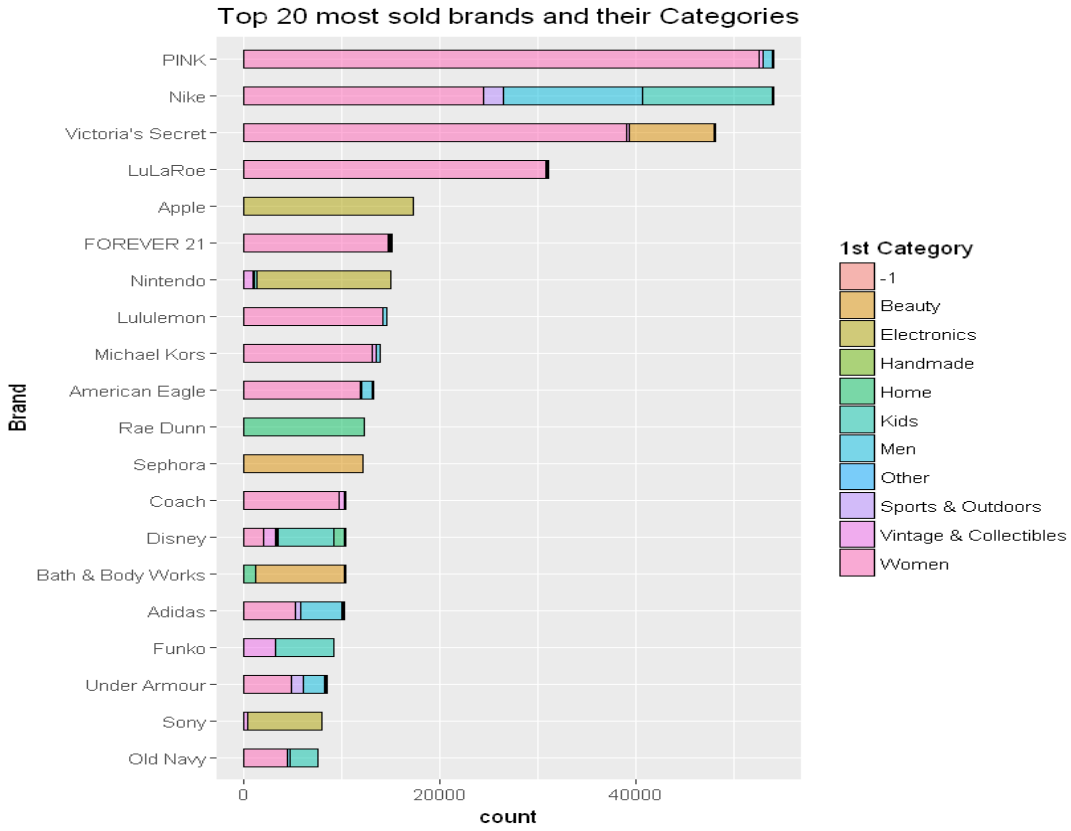


Fig 5.12

We can see that most of the brands such as "PINK","VICTORIAS SECRET", "NIKE" and "LULA ROE" are mostly sold by women.

- **Names:**

Analysis of top 20 names mostly used

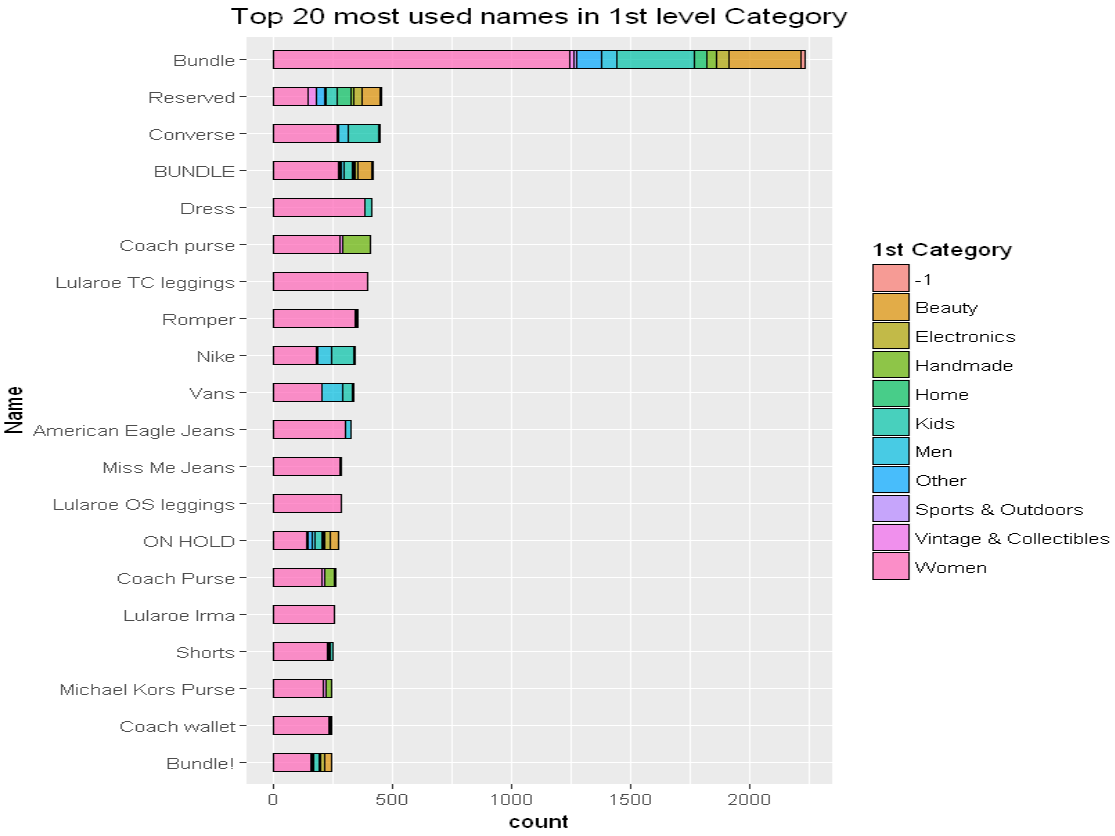


Fig 5.13

So clearly this gives us an insight of what 'name' consists of how this feature can be used in modelling.

Now we are half way around and we are still clueless of these affect prices as we cannot see a good relationship with the target.

Now let's proceed with Bivariate analysis of character class variables with target variable.

Analysis of top 30 items by median price:

Fig 5.14

Analysis of top 30 brand's by median price:

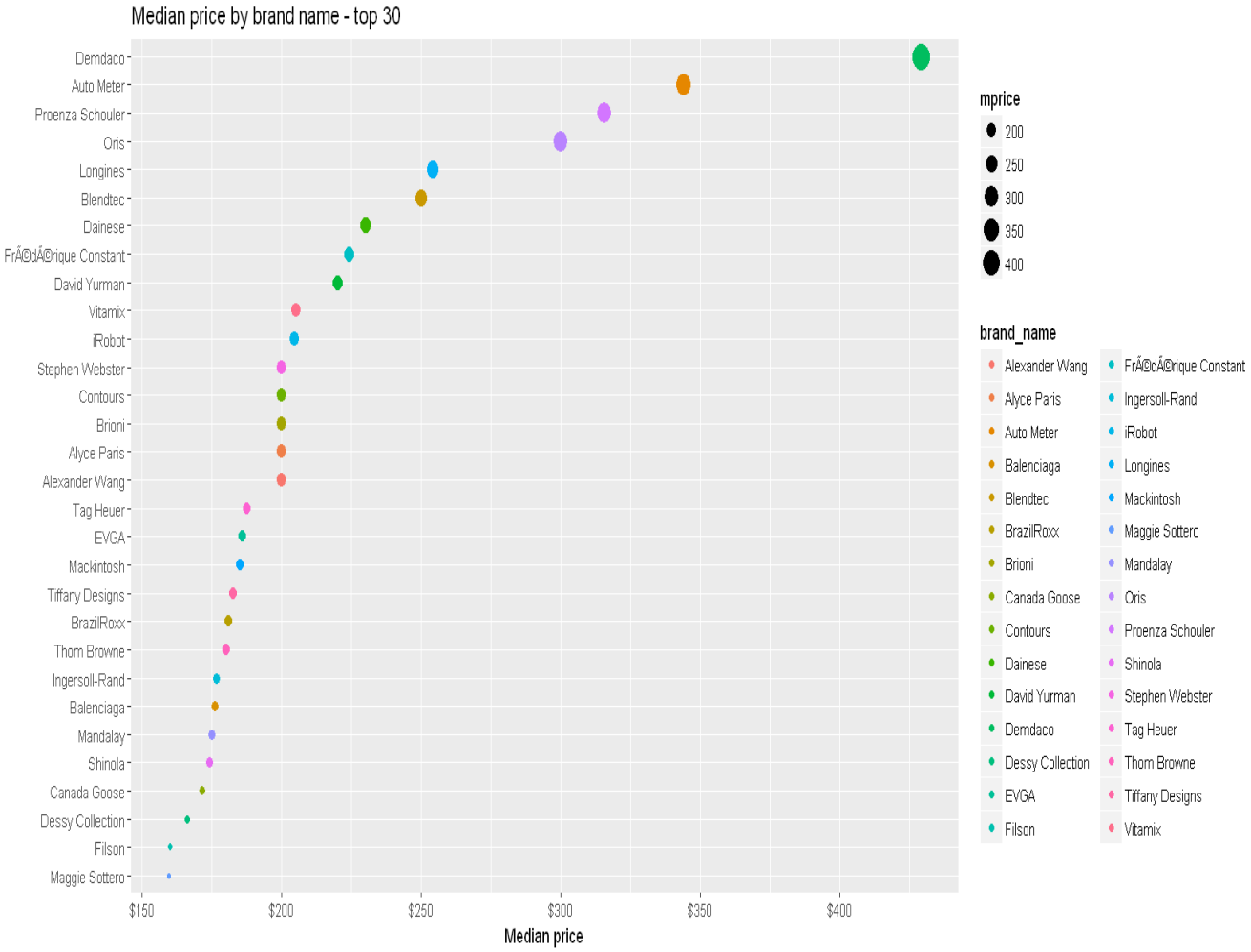


Fig 5.15

As seen from above we saw how cateogries, name, brand name influence price. categories, brand name and name are contributing so much into the response variable 'price'.

Now since we know that major information should be in "category_name", "brand_name", "item_description" "name" of the item. So, we shall do feature engineering to extract a few variables from these three variables.

Now let's analyze the '0' priced items, as these items can significantly have an impact on our model.

There were 874 rows with zero priced items which is a small amount when compared to the dataset, so let's drop them.

Combining the train and test data set for item description analysis.

6.0 Text Processing and Exploration

For the features containing text information such as item_description and name word clouds were formed to analyze the content .

Wordcloud for item_description

Fig 6.1

Wordcloud for name.

Fig 6.2

7.0 Feature Engineering and Data preparation for modelling

It was clearly evident that the text data has significant contribution towars price from Fig 5.14 and 5.15. Based on the above treemap's and wordcloud's the most significant contributor's were analyzed and 50 new features from name, brand_name, categories, and item_description were extracted.

```
beauty,electronic,handmade,home,kid,women,men,athletic,wine,athletic_apparel,
makeup,blouse,shoes,toy,,jewelry,phone,bag,dress,pant,accessories,luxury,tshi
rt,face,game,lip,eye,care,top,pink,secret,nike,apple,lularoe,size,color,leath
er,new,len_des,cond,tags,no_tags,bundle,coach,americaneagle,missme,
reserved,onhold,michaekors,converse
```

Overview of 50 new features extracted

Conversion of the character and factor variables to numeric

The length of unique set of characters from feature were taken and saved and the length of unique set of characters were assigned to the total levels of the particular feature and then convertig them to numeric.

Based on the data given the data was split between train and test.

8.0 Data Modelling

The model used here is Xtreme Gradient Boosting.

Boosting is a sequential process; i.e., trees are grown using the information from a previously grown tree one after the other. This process slowly learns from data and tries to improve its prediction in subsequent iterations.

The basic idea behind boosting algorithms is that the model capitalizes on the misclassification/error of previous model and tries to reduce it.XGBoost also uses a higher-order approximation, hence learns a better tree structure.

Coming to high dimensional problem like this, tree boosting beats the curse of dimensionality by not relying on any distance metric. Also the similarity between data points are learnt from the data through adaptive adjustment of neighbourhoods. This makes the model immune to the curse of dimensionality.

Also deeper trees help to capture the interaction of the features. Thus there will be no need to search for appropriate transformations.Thus, with the benefits of boosting tree models, i.e. adaptively determined neighbourhoods, XGBoost in general should make a better fit than other methods. They are able to perform automatic feature selection and capture high-order interactions without breaking down.

XGBoost sets a T_{max} and regularization parameter to make the tree deeper while still keeping the variance lower. The Newton boosting used by XGBoost is likely to learn better structures compared to other gradient boosting. XGBoost also includes an extra randomization parameter, i.e. column subsampling, this help to reduce the correlation of each tree even further.

Now I firmly believe that explains why we are using XGboost here.

Now let's have an overview of Newton tree booster used by XGBoost

Input : Data set $\mathcal{D}$.
A loss function L.
The number of iterations M.
The learning rate η.
The number of terminal nodes T

Initialize $\hat{f}^{(0)}(x) = \hat{f}_0(x) = \hat{\theta}_0 = \arg\min_{\theta} \sum_{i=1}^{n} L(y_i, \theta)$;

for $m = 1,2,..,M$ **do**

$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}^{(m-1)}(x)}$;

$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}^{(m-1)}(x)}$;

Determine the structure $\{\hat{R}_{jm}\}_{j=1}^{T}$ by selecting splits which maximize $Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} - \frac{G_{jm}^2}{H_{jm}}\right]$;

Determine the leaf weights $\{\hat{w}_{jm}\}_{j=1}^{T}$ for the learnt structure by $\hat{w}_{jm} = -\frac{G_{jm}}{H_{jm}}$;

$\hat{f}_m(x) = \eta \sum_{j=1}^{T} \hat{w}_{jm} \mathrm{I}(x \in \hat{R}_{jm})$;

$\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x)$;

end

Output: $\hat{f}(x) \equiv \hat{f}^{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$

Cross Validation was performed with 'gbtree' booster for 5 k- folds and 700 rounds with a learning rate of 0.3

The best iteration was found to be 700 itself.

Model was build on training data and tested on test data.

1st model:

The model was trained with a learning rate of 0.3 with same number of rounds.

Model Evaluation on training set

train-rmse:0.517222

train_rmsle:0.661780

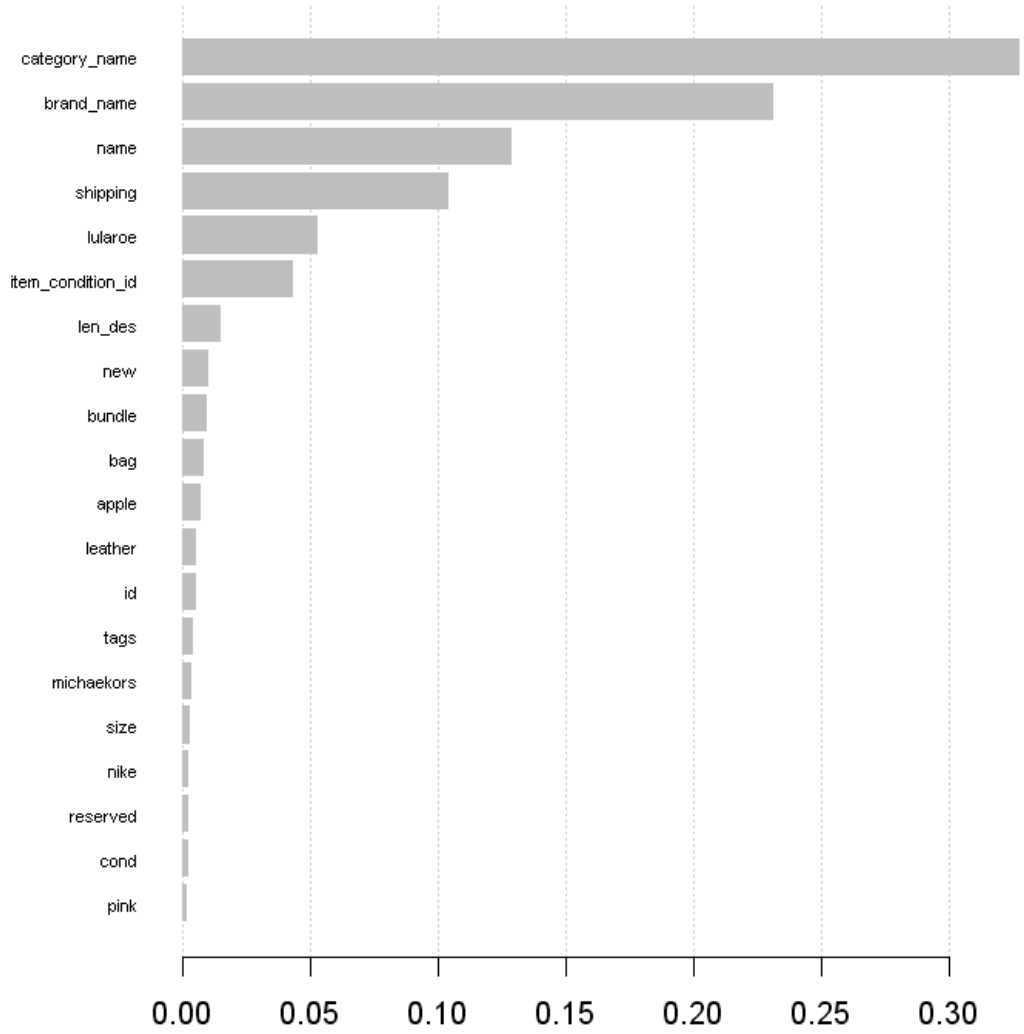
Model Evaluation on test set

test_rmsle:0.737781

Ok, so that was not great , let's consider the feature importance and tune our model.

Variable importance

To understand where we can improve, we first need to understand exactly which variables model liked and which it didn't.



This chart brings about an interesting point. In my opinion, this is what makes this project more excting, which means we have a tradeoff to consider: do you try and increase your predictions' accuracy by creating more variables, or by commanding more predictive rounds for your model? Like most things the answer is likely to be a balanced approach.

Though many of the variables we created are contributing to our model, we are going to begin removing some of the non-predictive variables and upping the rounds for our xgboost model.

We select all of the first 20 features which contribute to the model as shown above and train and test our model.

2nd Model:

Model Evaluation on training set

train-rmse: 0.518891

train_rmsle:0.32477

Model Evaluation on test set

test_rmsle:0.446741

Ok, so that's not a tremendous improvement but comparatively good to previous, so now let's tu ne and up the rounds and see how our model performs.

Model tuning:

Model was tuned with an increase in nrounds to 1200 and a learing rate of 0.7 and validated.

3rd Model:

Model Evaluation on training set

train-rmse:0.515994

train-rmsle:0.325167

Model Evaluation on test set

test-rmsle:0.448090

Note: The 3rd model was also tuned with parameter gamma=2 to prevent overfitting and modele d with the above parameters with a early stopping rounds as 20 and the model retuned early sto pping at 637th iteration meaning the model didn't improve from the past 20 rounds this was co mparatively similar to the 2nd model.

Hence by comparing the 3rd with 2nd , 2nd was found to be efficient in predicting the price.

From the above it is evdent that tuning with increase in learning rate and nrounds was found to be inefficient as there was no drop in the error rate hence we'll freeze our model with 2nd model

Conclusion:

Based on the predicted price from three different models the 2nd model's price list has been provided.

_____THANK YOU_____