Vinh Vu

Mercari price prediction

Springboard

04/01/2019

I.Introduction:

Product pricing at scale is a difficult task, 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. Through this project, I want to build an algorithm which can suggests product prices for online retailers.

My client, Mercari, is one of Japan's biggest online shopping sites. In order to improve better serve their customers, Mercari wants to offer price suggestion to their sellers when they are posting a new product for sale on their site. Based on my prediction, Mercari would be able to:

- Identify the product detail such as category, brand and type.
- Offer a price suggestion based on products' detail.
- Modify the product price before selling on the Mercari marketplace.

II.Data obtaining and cleaning:

Mercari dataset which is provided by Kaggle, include train.csv, and test.csv. There are some missing values in the dataset as well and I imputed them with Nan's for simplicity. Here are some of the pre-processing steps I did:

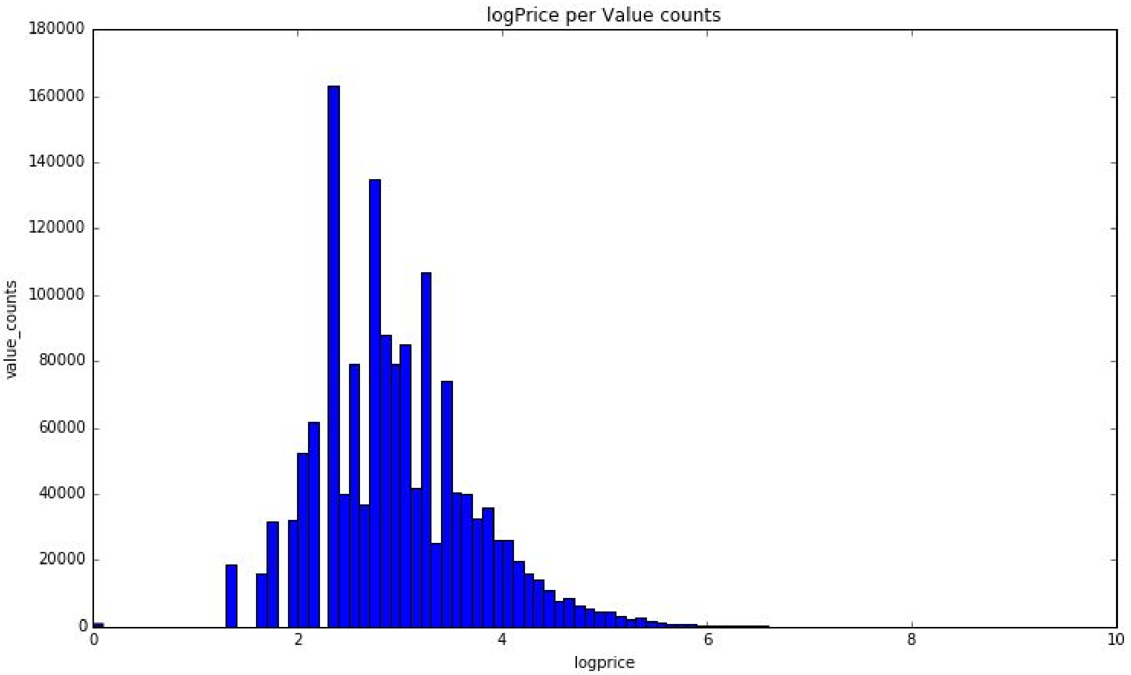
- Handling Missing Values – Replaced with Nan.
- Lemmatization performed on item description.
- Label Encoding – Turned categorical columns into 0's and 1's.
- Tokenization – Given a character sequence, tokenization is the task of chopping it up into pieces.
- Scaling – Scaled the price variable (log price).

Dataset:

Train_id / test_id
Name – name of product.
Item_condition_id – product's condition.
Category_name – product's category.
Brand_name – product's brand.
Price – price.
Shipping – shipping fee.
Item_description – product's detail.

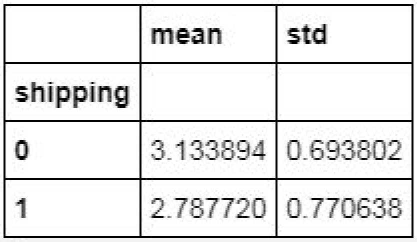
III.Data Story:

1.Price:

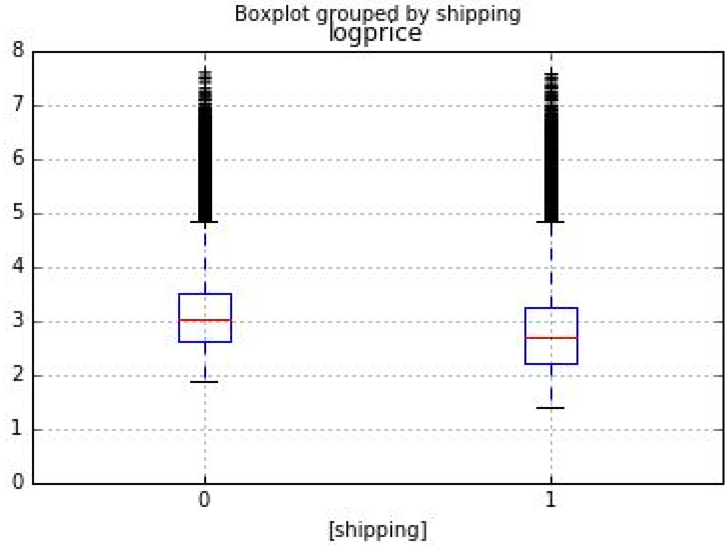


I create a log price scale column and histogram graph. in this project, the regular price scale shows an equal distance between prices, and each unit change on the chart is represented by the same vertical distance on the scale, regardless of what price level the asset is at the change occurs. However, the log price scale is plotted so that the prices in the scale don't have a same equal distance; instead the scale is plot in the way that two equal percent changes are plotted as the same vertical distance on the scale. This log price per value counts graph is skewed left .

2.Log Price with shipping:

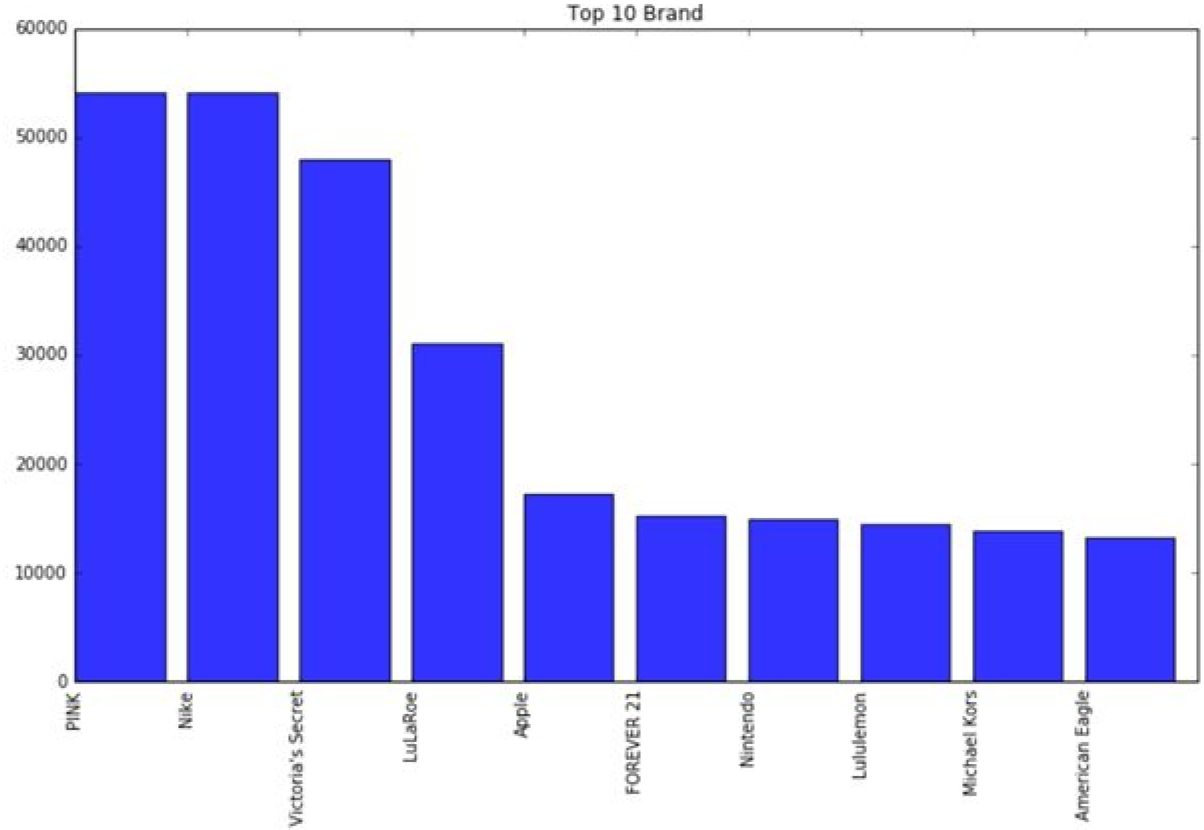


	mean	std
shipping		
0	3.133894	0.693802
1	2.787720	0.770638



The mean of log price without shipping is 3.13 and the mean of log price with shipping is 2.78 .

3.Top Brand Distribution:

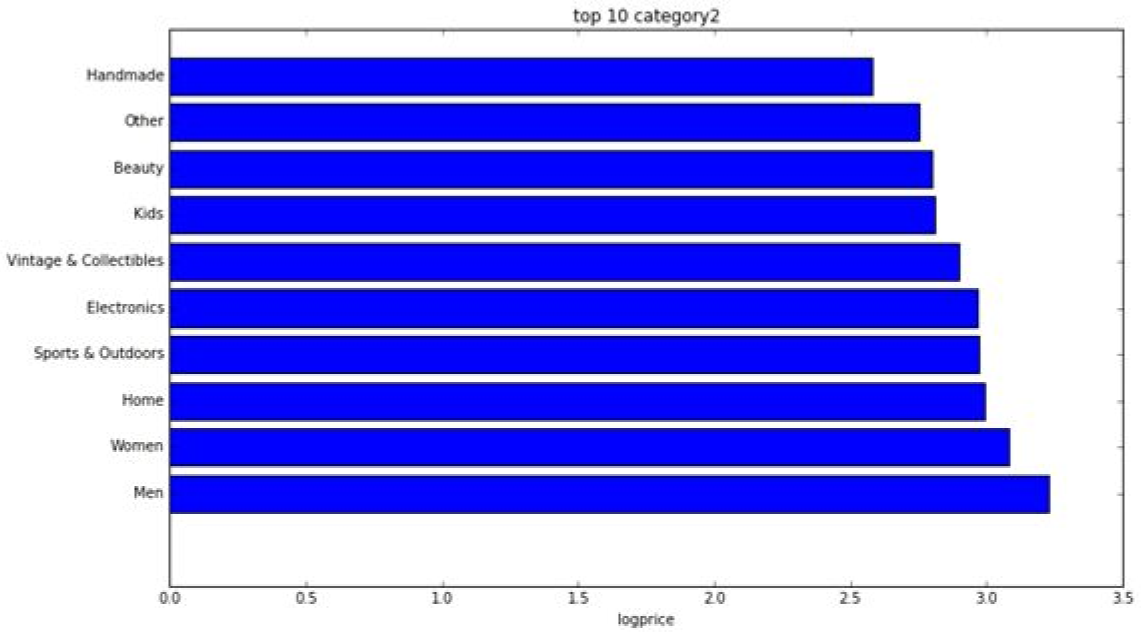


The majority of the top brands are clothing brands and electronics. Of the top 3 brands, two of them, Victoria’s Secret and Pink, are aimed exclusively at female customers.

4.Category:

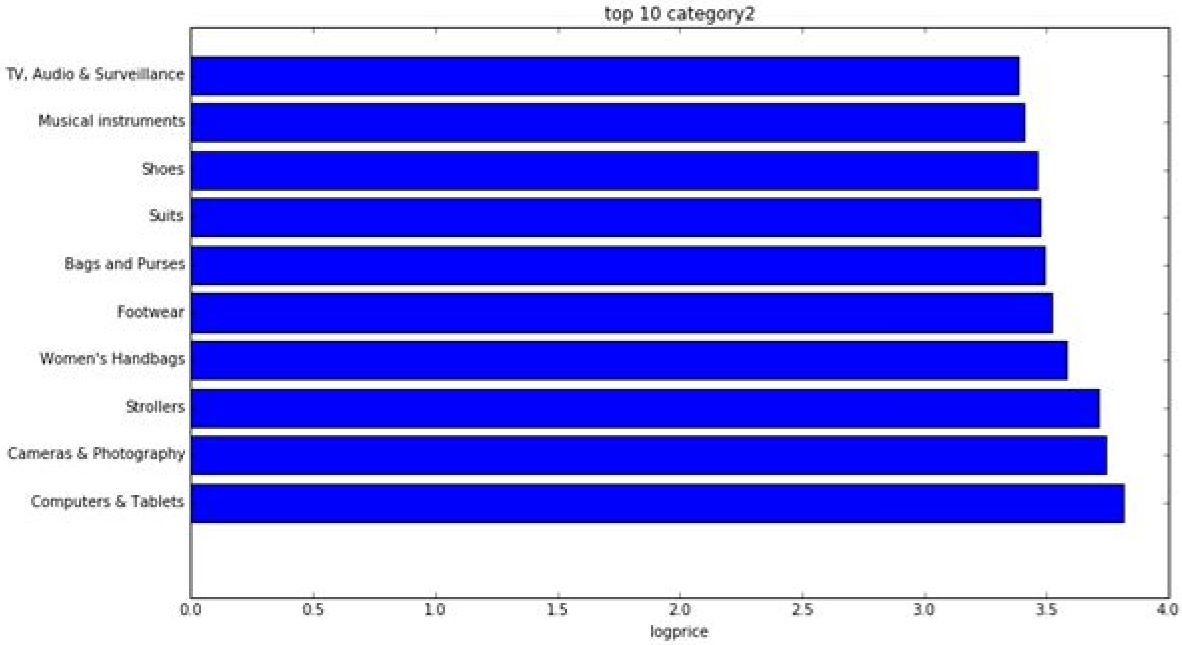
The data provide 3 categories for each item, generally each one following a sub-category of the previous category. I decided to split these into three different features: category 1, category 2 and category 3

Category 1:



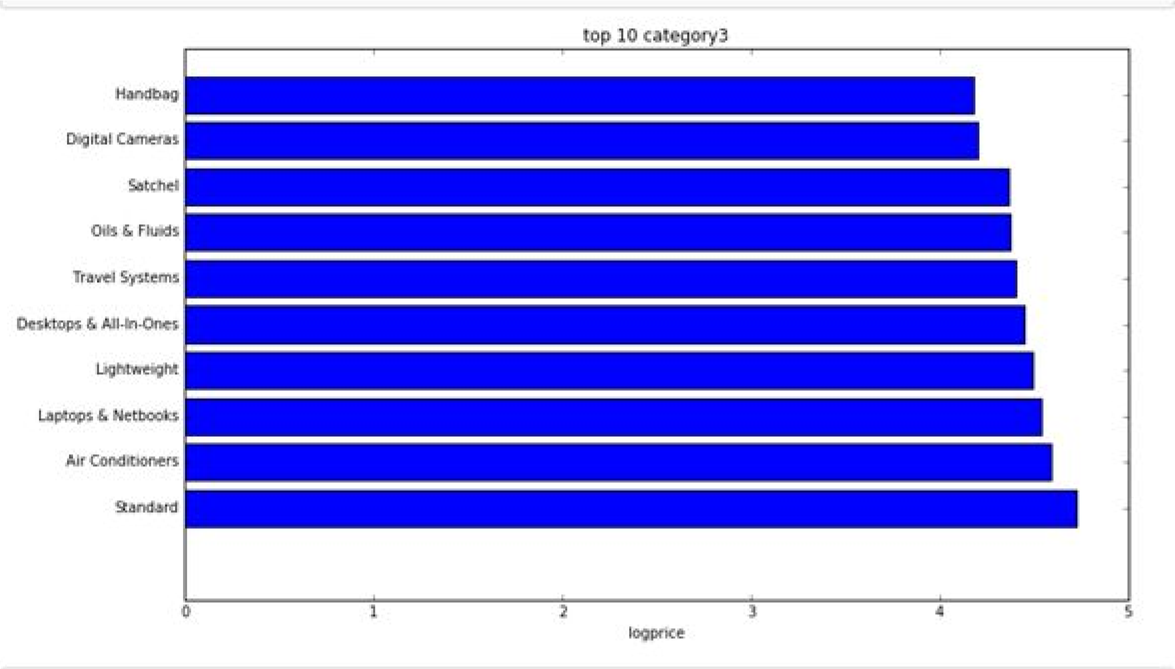
The top 3 most expensive categories are Men, Women and Home. The majority of the most expensive categories fall into clothing, home and electronics.

Category 2:



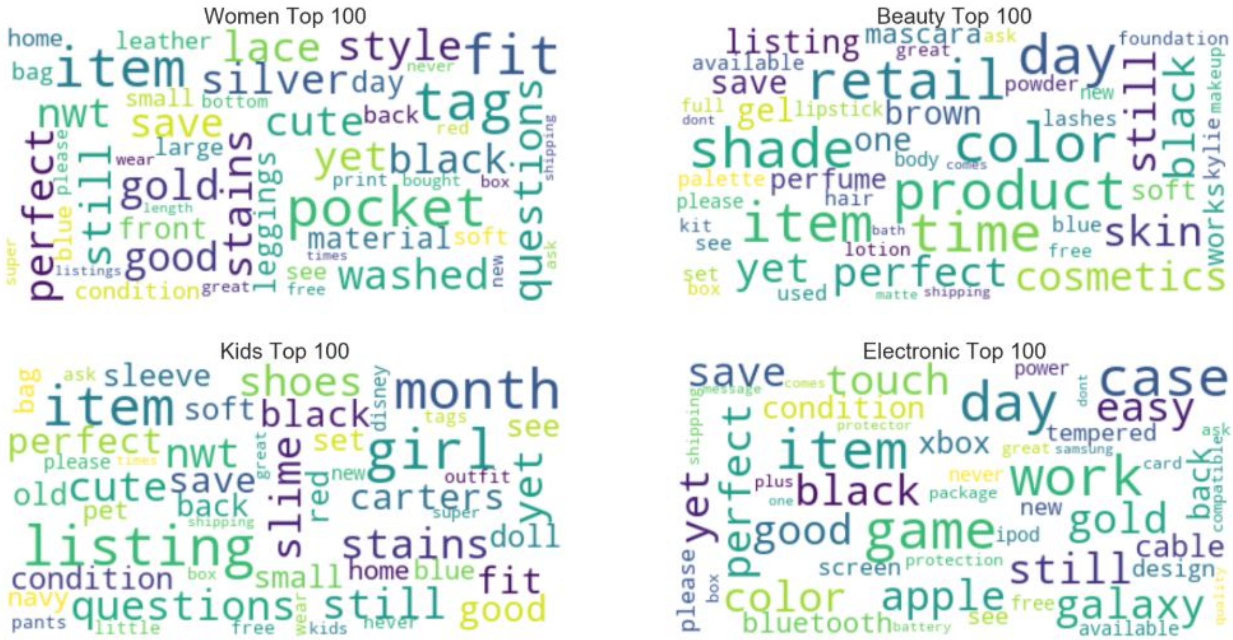
Majority of the top categories are electronics. Computer & tablets, camera & photography and strollers are in top 3 categories, but Computer & tablets and camera & photography fall into electronics.

Category 3:



Majority of the top categories are clothing and electronics. Standard, air conditioner and laptop & netbooks are in top 3 categories, but air conditioner and laptop & netbooks also fall into electronics.

5.Item Description by Category:



Across all the categories, item description commonly contained the words shipping, color, and perfect. This could indicate a few things. “Shipping” may indicate that customers may be looking for free or low cost shipping, “color” may indicate that customers are looking for multiple color options, or that color is just an important feature for products generally, “perfect” could indicate that customers appreciate products on perfect condition, or that the products that customers like perfectly meet their needs.

For the women’s clothes category, the common words in item description are tag, fit and pocket. The “tag” word appears in 13% in total women product, and “fit” appears in item description in 11% of all product. Therefore, it may indicate that women are looking the clothes that fit and having tags in perfect condition. Moreover, the word “pocket” appears in 3% of all women’s clothes product, but it is mostly in woman bottom products such as legging, jean, sportswear and skirt.

For top 100 beauty, the commonly words in item description are color, shade and retail. The “color” word appears 16% in total beauty product, it is mostly in makeup and nail products. Also, the word “shade” appears 8% in total beauty products, it is mostly in makeup products. Moreover,the “retail” word appears s 6% in total beauty product, the sellers shows the retail price in the item description. Therefore, it may indicate that customer may consider the product’s price which compares to retail price, color and shade options for makeup and nail products.

For top 100 kids, the most common words in item description are girl, month and listings. The “girl” word appears 9% in total kid product. The “month” word appears 8% in total kid product. It may indicate that customers may be looking for baby size by month and girl clothes. Moreover, the word “listings” appears 2% in total kid product, the sellers advertise for their other products in listings.

For electronic category, the most common words in item description are case, work and black. “Case” may be just a very common item as cases required for many different types of electronic items, and “black” may indicate the most common color. The “work” indicates that the ease of functionality for electronic items are of paramount importance.

6.Item description rate:

word vs rate:

In order to determine expensive and cheap words, I used the first 1000 rows from each of four different categories: woman, kids, electronic and beauty and I applied naive bayes machine learning.

Pre-Processing: create a new column “rate” which compare between log prices. If log price is bigger than log price mean, we label it 1, otherwise it is labelled 0.

Naive Bayes method:

The probability model that was formulated by Thomas Bayes (1701-1761) is quite simple yet powerful; it can be written down in math equation:

For a document *d* and a class *c*

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Apply Naive Bayes algorithm for this project :

1. Iterate over the labelled rate word in item description and, for each word w in the item description, count how many of the expensive rate contain w. Compute

$$\text{P(w | E)} = \frac{|\text{expensive rate cointaing w}|+1}{|\mathit{expensive\ rate}\ |+2}$$

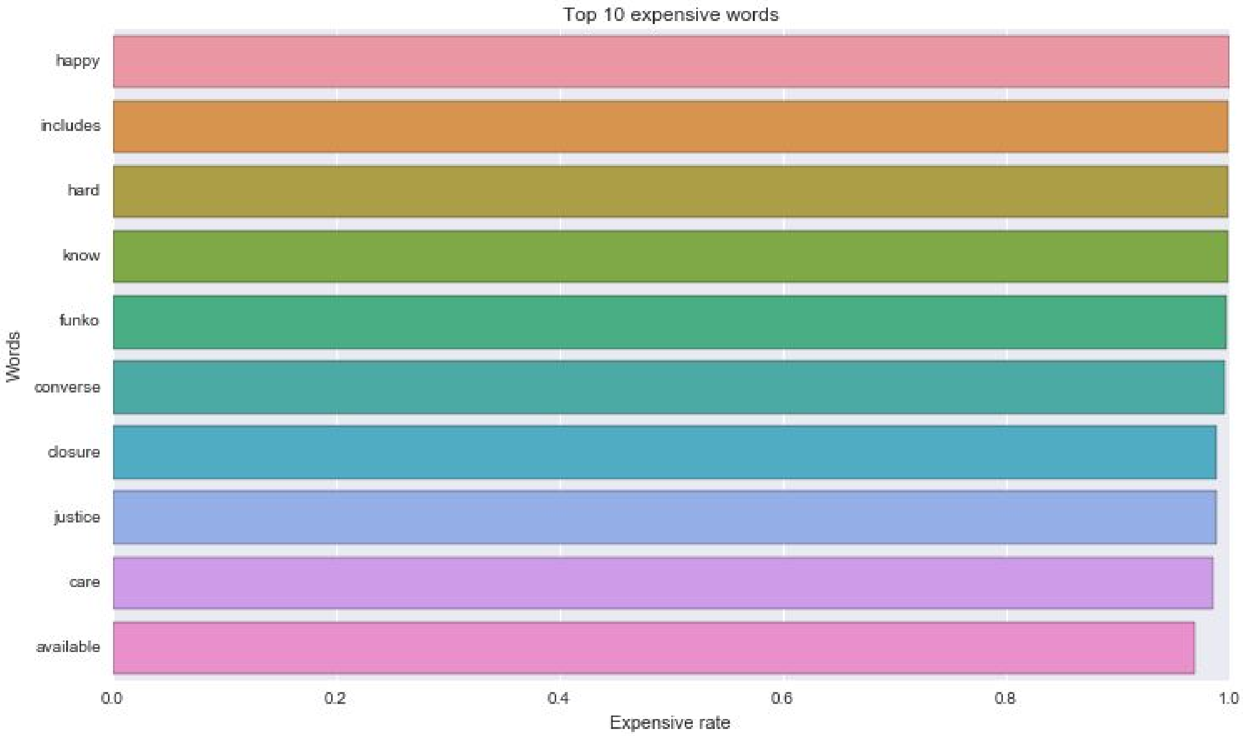
2. Compute P (w | C) the same way for cheap rate.
3. Compute P(E) = |expensive rate|/ (|expensive rate |+| cheap rate |)
4. P(C) = |cheap rate|/ (|expensive rate |+| cheap rate |)
5. Given a set of unlabeled test item descriptions, iterate over each:
   a. Create a set {x1, . . ., xn} of the distinct words in the item description. Ignore the words that you haven't seen in the labelled training data.
   b. Compute

$$P(E|x1, \dots, xn) = \frac{\text{P(E)} \prod P(xi|E)}{\text{P}(E) \prod P(xi|E) + \text{P(C)} \prod P(xi|C)}$$

   c. If P(E|x1,…,xn)>0.5 output "1", else output "0"

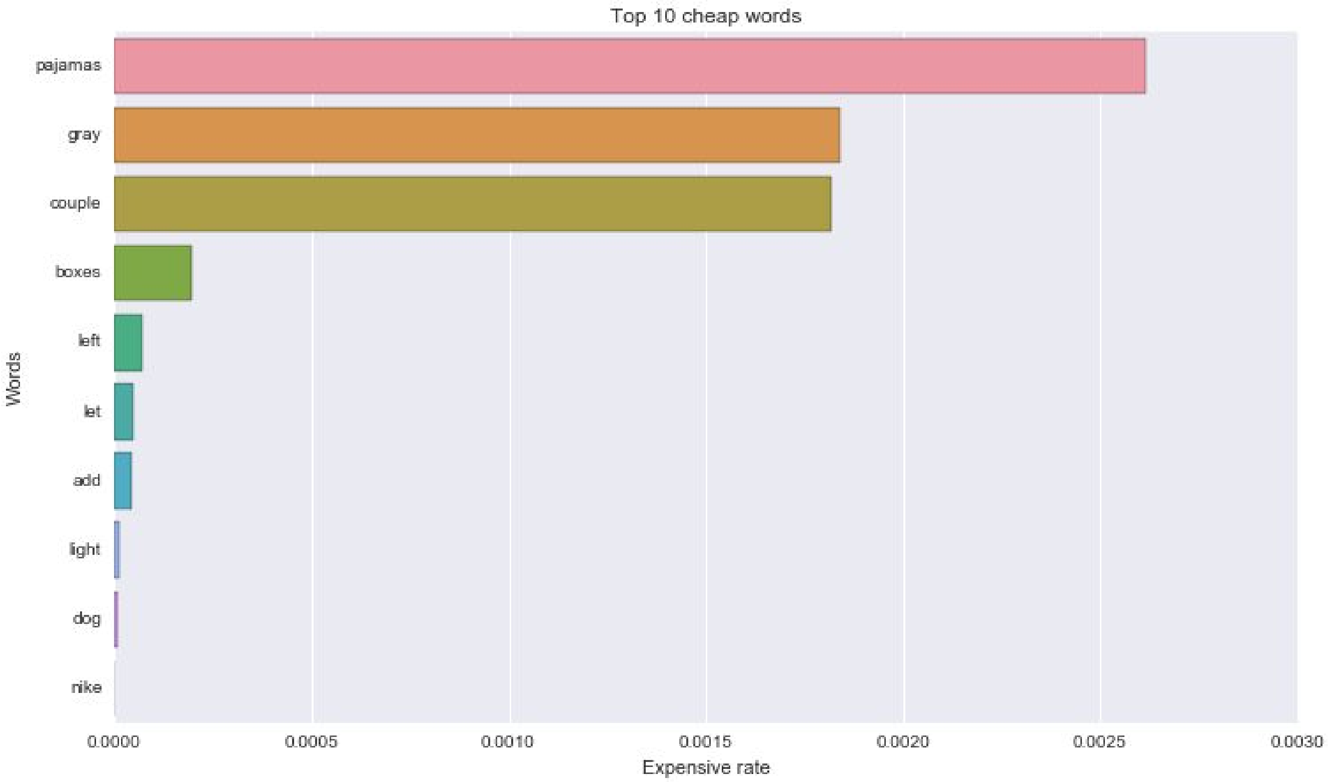
Expensive Words for "Woman" Category:

Across expensive words in woman category, item description commonly contains words like hours, light and bling. The "hours" word indicates for shipping time such as within 24 hours. In addition, the "light" word prefers for product's material or color . The " bling" word is used as logo mostly in victoria secret products. For all top words, they are used 100% of the time in expensive products which is counterintuitive and may be a result of making 'expensive' half of the dataset. All words appear more than 50 times so this is not solely due to small data size.



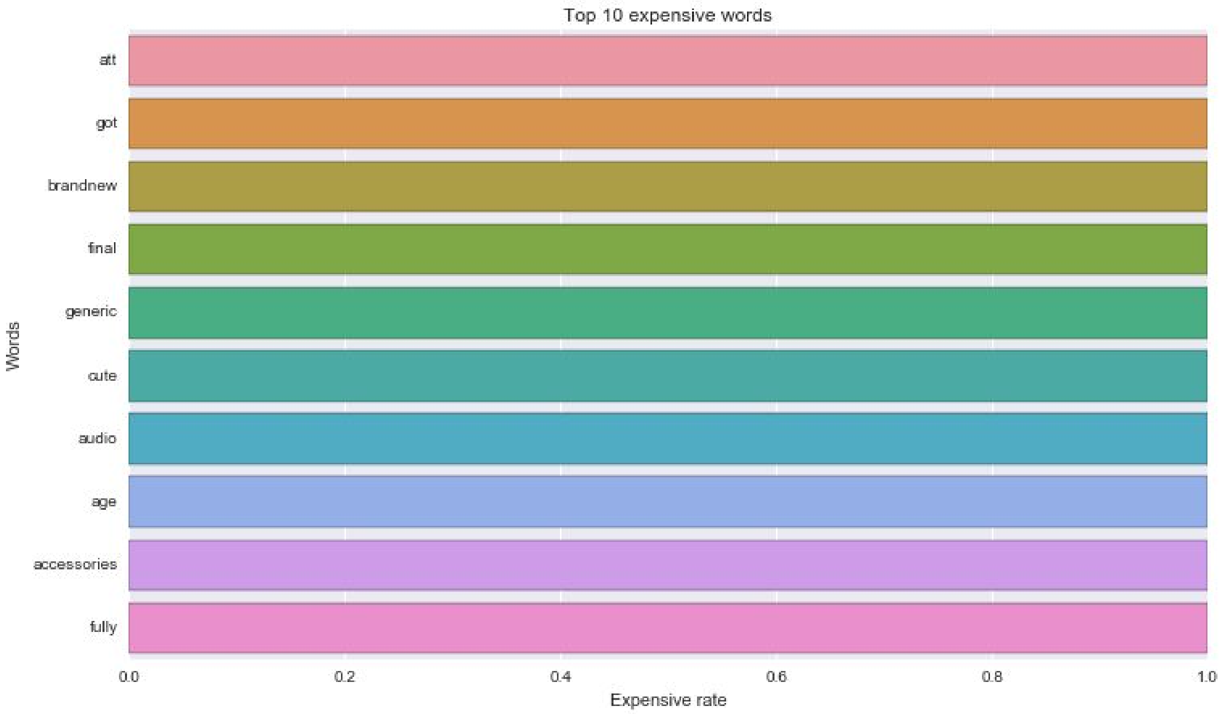
Cheap Words for "Woman" Category:

Across Cheap words in woman category, item description commonly contains words like hlf, multiple and color which are mostly in items with cheap rate . The "color" word is used 9% in first 1000 woman products to describe the color products. The "multiple" word is used 0.5%in first 1000 woman products to stand for various size and color of product offers. The "htf" word is also used 0.5% in first 1000 woman products which is a brand of LuLaRoe.



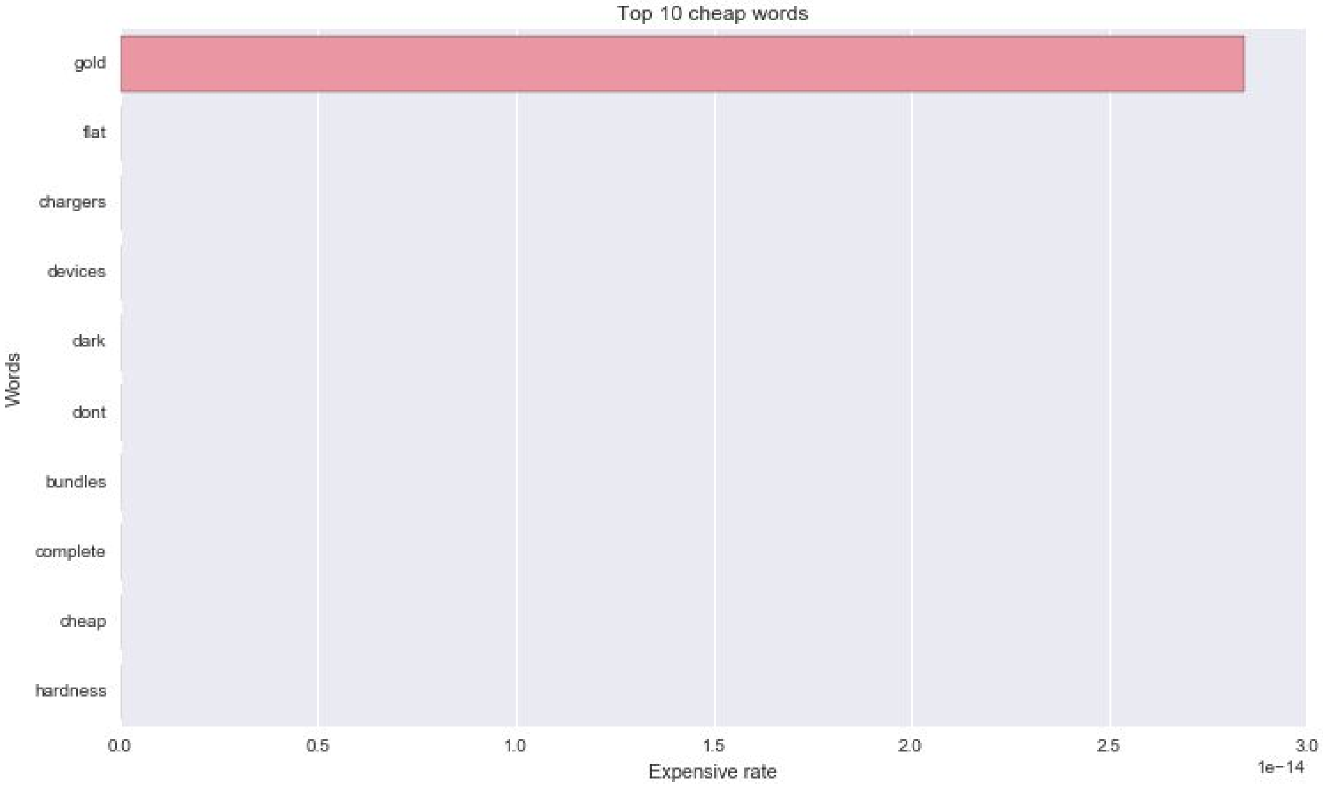
Expensive Words for "Electronic" Category:

Through the electric category, the commonly words are att, got and brandnew. The "att " word stands for a phone from AT&T telecom or battery type. The "got" is used mostly to describe a reason to get product from seller and product is mostly reused. The" brandnew" is used for products's condition. For expensive words, they are used 100% of the time in all expensive items.



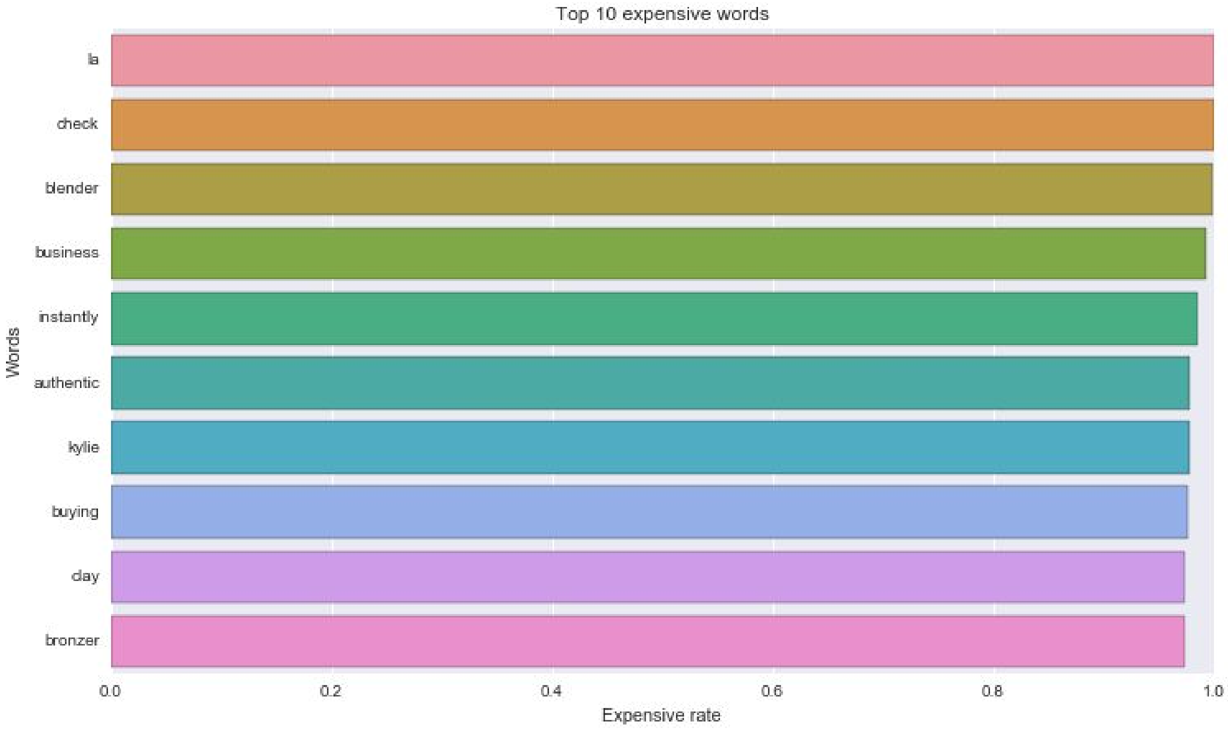
Cheap Words for “Electronic” Category:

Across Cheap words, item description commonly contains words like hardness, cheap and complete which 100% appeared on cheap items; such as phone case and protection screen. The “hardness” word is used 0.9% in first 1000 electronic products to describe the phone protected screen which is very cheap. The “ cheap” word is used 2% in 1000 electronic products which is mostly appeared for phone cases. The “ complete” word is used 2.3% in 1000 electronic products to describes about complete series product; such as complete all season Grey anatomy DVD, or complete all Resident Evil games.



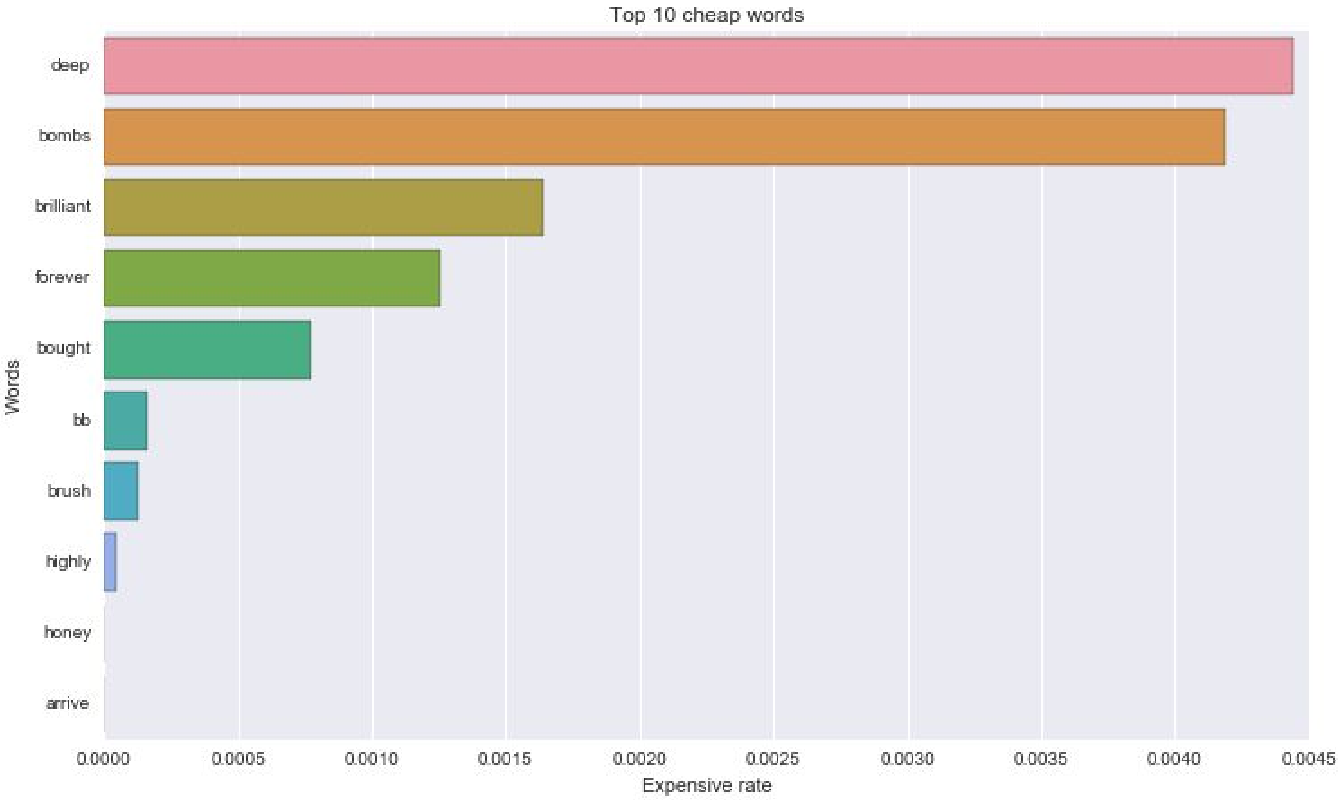
Expensive Words for “Beauty” Category:

Across expensive words, item description commonly contains words like la, check and blender. The “la” word is commonly used for eyelashes. Also, the “check” word is advertised for new product and ask customer to check it out. Finally, the “Blender” is stand for Beauty Blender brand. “La”, “check” and “blender” are 100% in expensive items .



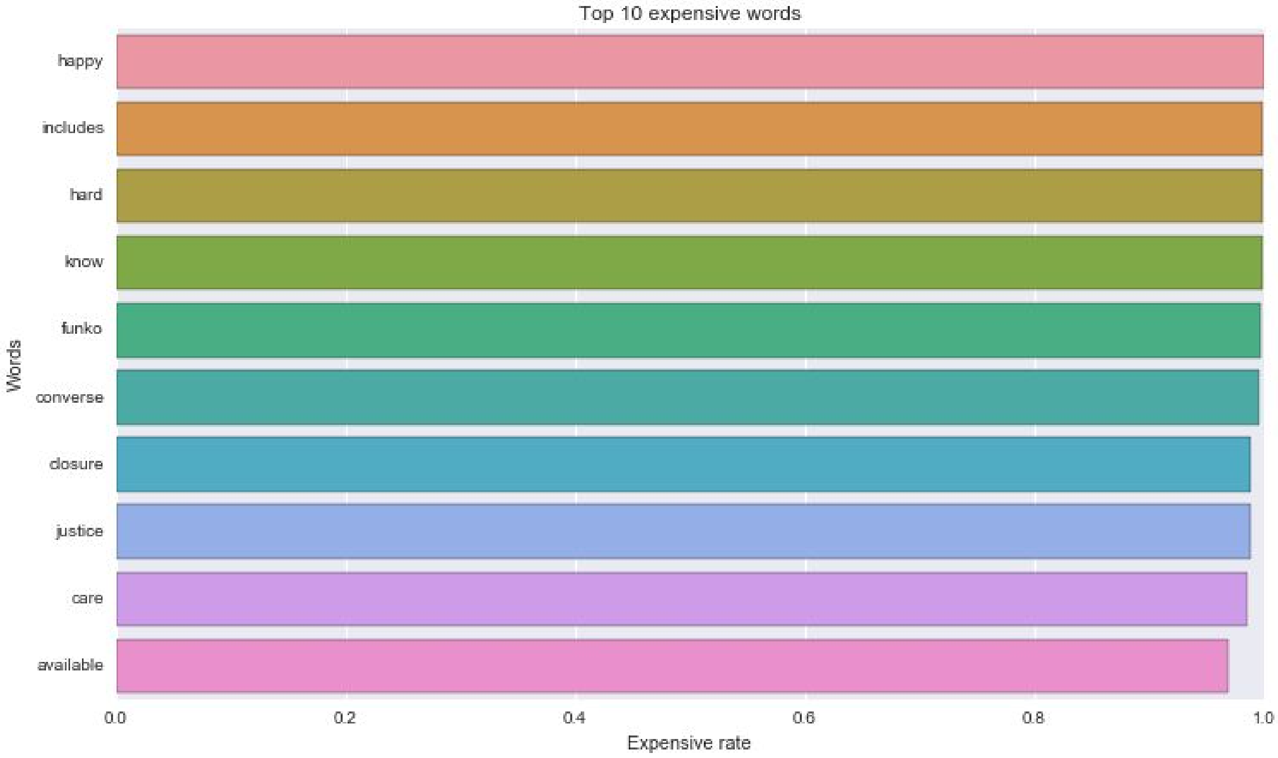
Cheap Words for “Beauty” Category:

Across Cheap words, item description commonly contains words like arrive, honey and highly which are more 99% to appear in cheap items . The “arrive”, repeats 1.2% in first 1000 beauty product, which is consider to value the product which is related to shipping and availability of product. The “honey” word is used 1.1% to describe bronzed shades of makeup products. The “highly” word is used 1.1% to stand for “ highly recommended”, but it doesn't guarantee that actual products treat all customers with the same result.



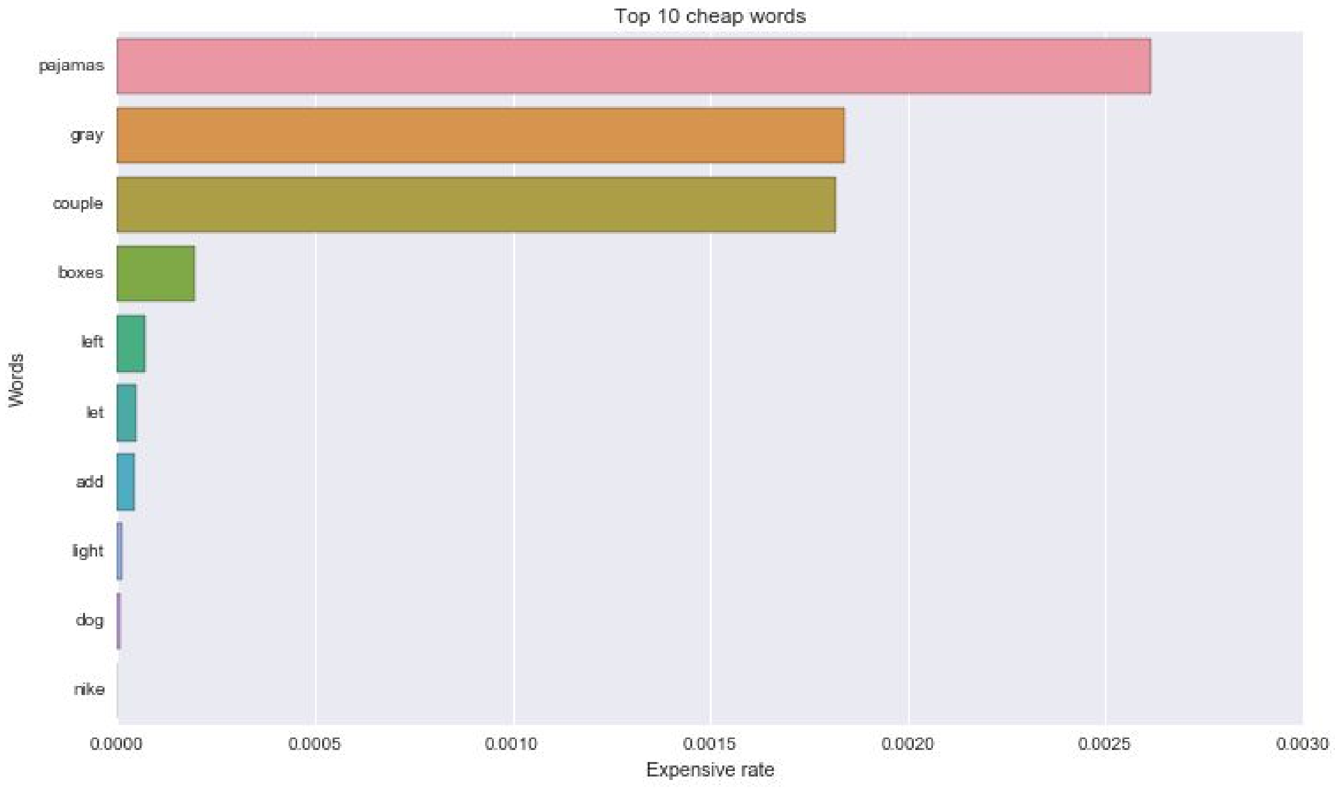
Expensive Words for "Kid Category:

Across expensive words, item description commonly contains words like happy , includes and hard, which used 100% in expensive items The " happy" word is described a bundle or clothes from various expensive brands like Ralph Lauren, Nike and Mam Mama. The" includes" word is described a gift or item that includes in the product. The " hard" word is describes the strength of toys.



Cheap Words for “Kids” Category:

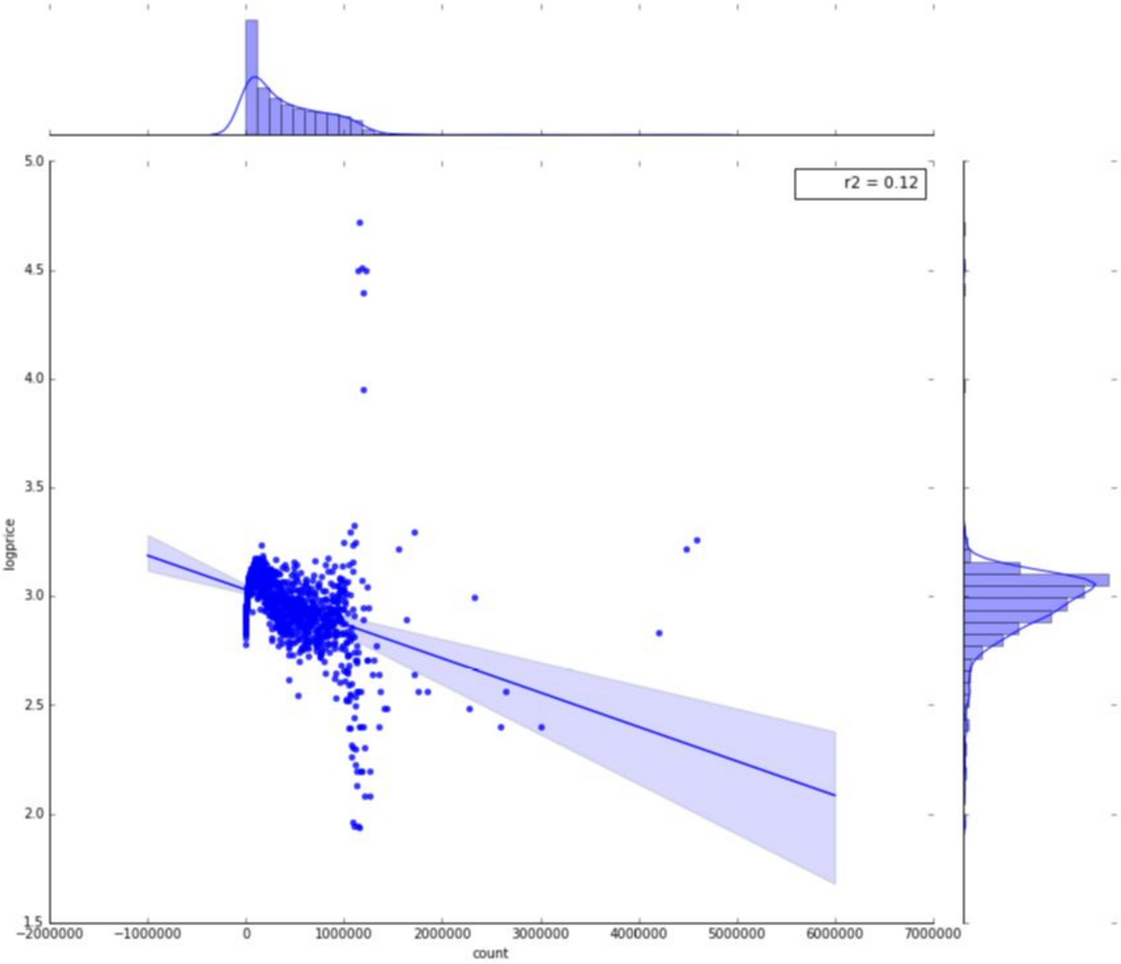
Across Cheap words, item description commonly contains words like nike, dog and light, which appears 99% in cheap items. The “nike” word is a NIke brand , mostly shoes in 3.3% of first 1000 kid products. The “dog” word is 1% and mostly toys. The “light” is 4.4% to describe the clothes material. For cheap words, they are mostly in cheap items



7.Word Count vs Log Price:

There is a positive linear relationship between word count and log price from about 0-300 words. After that there is a gradual negative relationship, which drops at about the 1000 word point. Because of this, I went ahead and created a new quadratic feature based on word count which has a more linear relationship with log price.

Quadratic Term of Word Count vs Log Price:



IV.Text Processing:

Text is an unstructured form of data and must be preprocessed. The main purpose of preprocessing data is cleaning, normalizing the text, and standardizing the data.

Terminology:

- Document: full text sentence or paragraph.
- Tokens: word features.
- Corpus: A collection of documents.
- Term Frequency (TF): How often a token is in a single document.
- Inverse Document Frequency (IDF): Distribution of a token over a corpus.

Pre-Processing Techniques:

1. Stop Word Removal: Removing common words with little predictive power such as “the”, “a”, “an”, “to”.
2. Bag of Words Representation: treats each token (words and n-grams) as a feature in the document, with term frequency as its value.

3. TFIDF (Term Frequency Inverse Document Frequency): Rather than using term frequency as the value as in a standard Bag of Words, dividing term frequency by inverse document frequency to . boost tokens that have low frequencies. For example, if the word “play” is common, then there is little to no importance. But if the word “mercari” is rare, then it has more weight/importance.
4. N-Grams: Sequences of adjacent words as tokens. For example, since a word by itself may have little to no value, but if you were to put two words together and analyze it as a pair, then it might add more meaning.
5. Stemming/Lemmatization: converting a word into its common base form.

V.Machine Learning:

For machine learning, I use first 5000 rows from woman category.

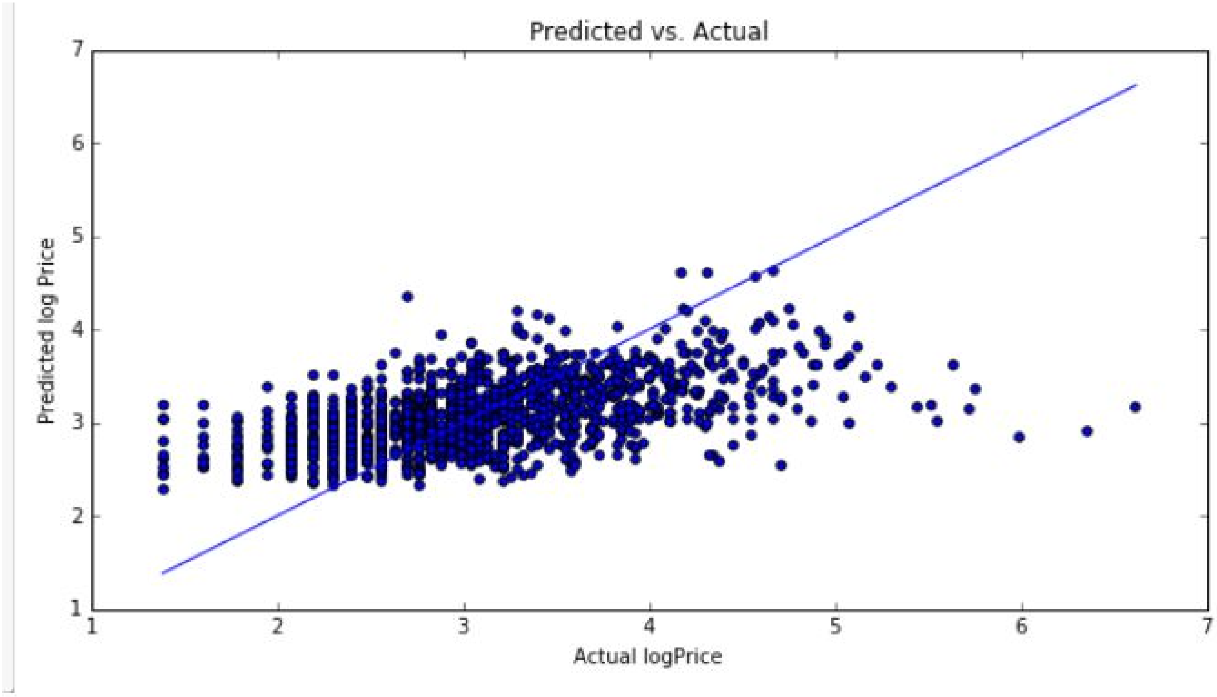
1.Random forest :

Pre-Processing data:

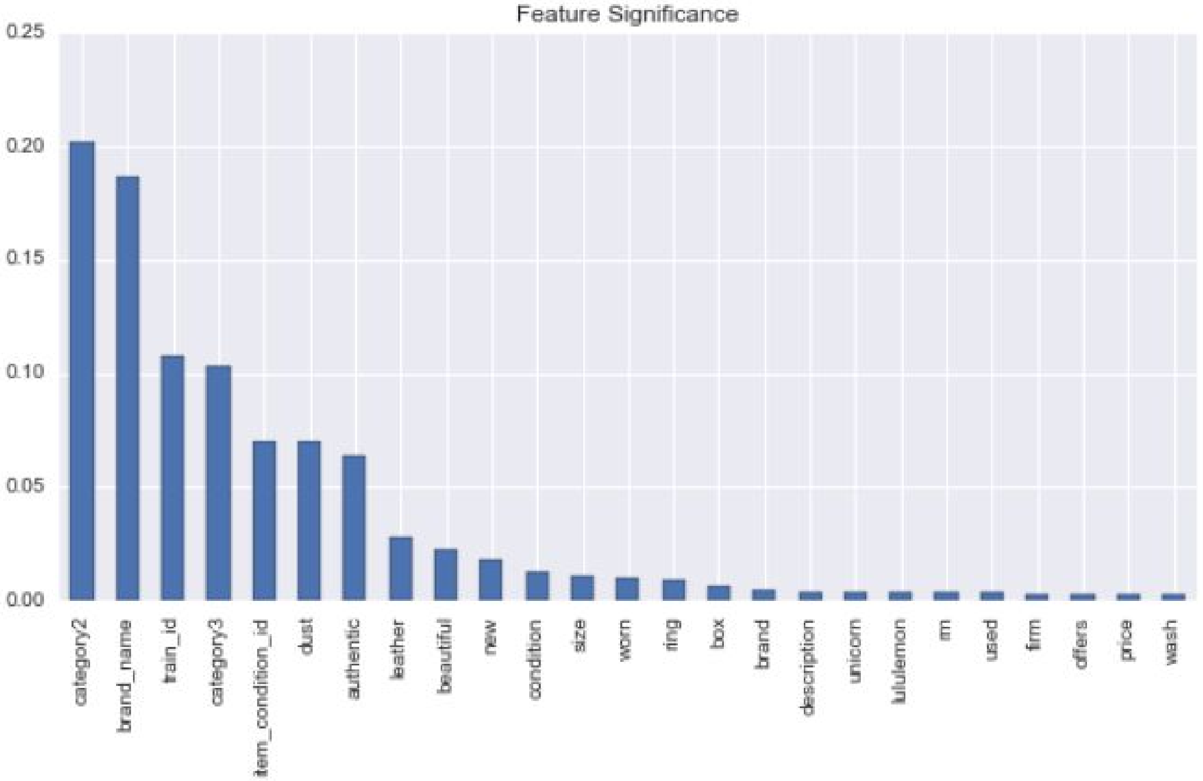
- TfidfVectorizer for name and item description.
- Split category into three categories, and categorize them.
- Categorize brand name.

Processing data:

- Using gridsearch to tune hyperparameter on n_estimators,min_samples_leaf and max_features.
- R^2 for training set: 0.49.
- R^2 score for testing set: 0.29.
- Cross validation RMSE: 0.63.



- Important features:



- Top three important features: category, brand name and train ID.

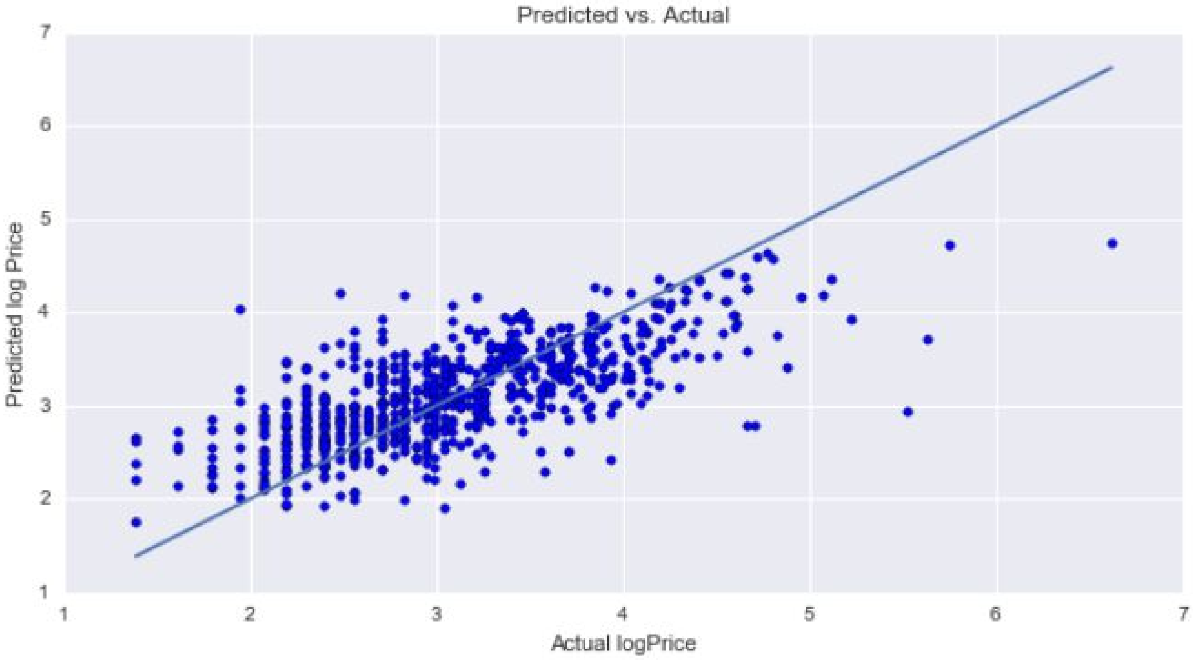
2. Ridge regression:

Pre-Processing data:

- TfidfVectorizer for name and item description.
- Split category into three categories, and categorizes them.
- Categorizes brand name.

Processing data:

- Using gridsearch to tune hyperparameter on alpha, max_iter, tol and random_state.
- R^2 score for training set :0.70.
- R^2 score for testing set :0.49.
- Cross validation RMSE : 0.56.



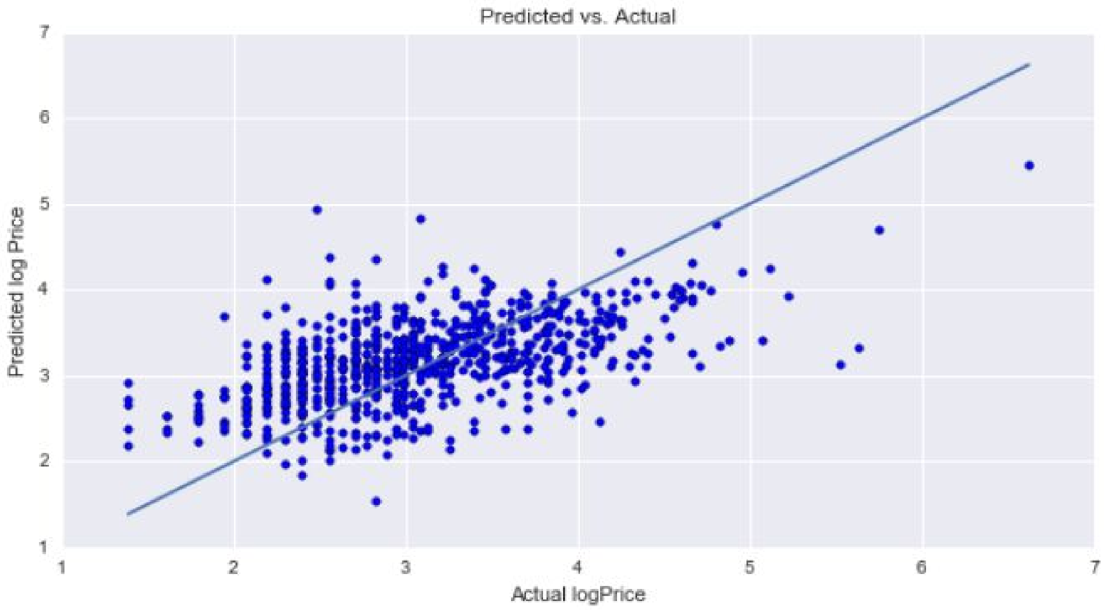
3. SVR:

Pre-Processing data:

- TfidfVectorizer for name and item description .
- Split category into three categories, and categorizes them .
- Categorizes brand name .

Processing data:

- Using gridsearch to tune hyperparameter on C, epsilon, gamma and kernel.
- R^2 score for training set :0.42.
- R^2 score for testing set :0.26.
- Cross validation RMSE : 0.65.



-

In order to determine the best machine learning method, I looked at the definitions of R square and RMSE. The R square is a measure that helps you to determine how well the model predicts responses for new observations. The R square training set should a little higher than the testing set in order to be a good approximation of the R square in the test set, but not too high. The RMSE indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. Lower values of RMSE indicate better fit. For this data set, the best machine learning is ridge regression.

VI. Conclusion:

This project has opened up my mind into the knowledge of NLP and it showed me how much pre-processing steps are involved to analyze text data. I learned the most common steps for text pre-processing and the choice of algorithms and how important computation is when you're dealing with large datasets. Through this project, I selected ridge regression was the best method in order to predict price. Mecarri can use my analysis in order to offer a suggestion price for the sellers by item detail and category . However, the weakness of this analysis is that it could not predict price for all items due to memory error, but I will try to learn other methods to optimize my analysis in future.