Evaluating Model Accuracy and Time Efficiency for Price Predictions on Online Retail Markets

I. Introduction

The e-commerce market is a massive and rapidly growing industry. An estimated 1.8 billion people made online purchases in 2018; and it is projected that, by 2040, 95% of purchases will be facilitated by e-commerce. Understanding the importance and prevalence of the online retail industry, we want to use machine learning models to evaluate a given item's conditions and features in order to make a price suggestion for sellers on an online second-hand shopping market. Product pricing is an important part of business operations. Prices can indicate the product conditions to buyers, and at the same time, signal the size of demand to sellers. It is especially relevant on the online second-hand markets where sellers are allowed to sell almost anything and the problem of asymmetric information is especially rampant.

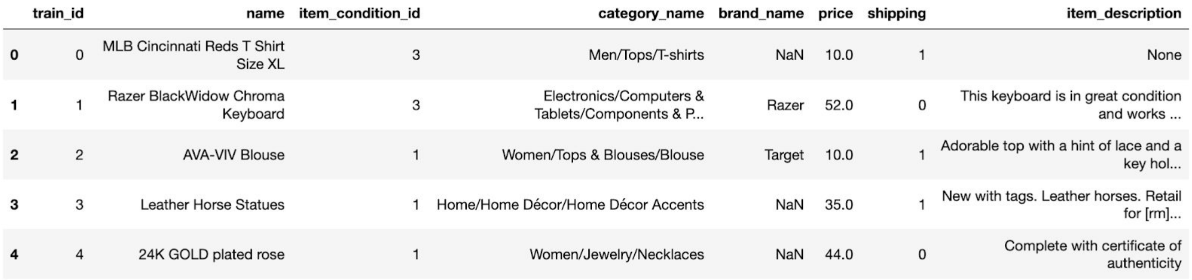
When making price predictions, small differences make huge impacts, especially considering a plethora array of products on online markets. For example, the prices for clothing and make-up products are heavily influenced by current fashion trends, and brand names; while electronics are influenced by product features and manufacturers.

Using various models, including linear regression, Lasso, KNN, and random forest, we were able to achieve an error rate (measured by MSE) at 0.6072. We found some variables are more important for price predictions than others. For example, whether free shipping is included plays an important role in predicting prices.

II. Methodology

Exploratory Data Analysis

The data was directly downloaded from Mercari Price Suggestion Competition on Kaggle. The dataset provides information on product names, item conditions, categories, prices, shipping, and item descriptions. Due to computational constraints, we used the first 10,000 observations for this project. The following table shows the first five observations in the dataset:



	train_id	name	item_condition_id	category_name	brand_name	price	shipping	item_description
0	0	MLB Cincinnati Reds T Shirt Size XL	3	Men/Tops/T-shirts	NaN	10.0	1	None
1	1	Razer BlackWidow Chroma Keyboard	3	Electronics/Computers & Tablets/Components & P...	Razer	52.0	0	This keyboard is in great condition and works ...
2	2	AVA-VIV Blouse	1	Women/Tops & Blouses/Blouse	Target	10.0	1	Adorable top with a hint of lace and a key hol...
3	3	Leather Horse Statues	1	Home/Home Décor/Home Décor Accents	NaN	35.0	1	New with tags. Leather horses. Retail for [rm]...
4	4	24K GOLD plated rose	1	Women/Jewelry/Necklaces	NaN	44.0	0	Complete with certificate of authenticity

Figure 1 Data description

For the purpose of exploratory data analysis, we performed Principal Component Analysis (PCA) on the item condition, category, brand names, and shipping. The plot below shows the data on two principal components. From the plot, there are mainly two clusterings. They are clustered by whether free shipping is included. This implies that free shipping is an important variable. It is worth noting that PC 1 explains 12.9% variance in the data and PC2 explains 7.75% variance. The total variance explained by the two principal components is relatively small. One potential reason for this is that all of the variables we used in PCA are categorical and PCA works better on continuous data[1]. Nevertheless, the plot still provides useful information on the importance of shipping.

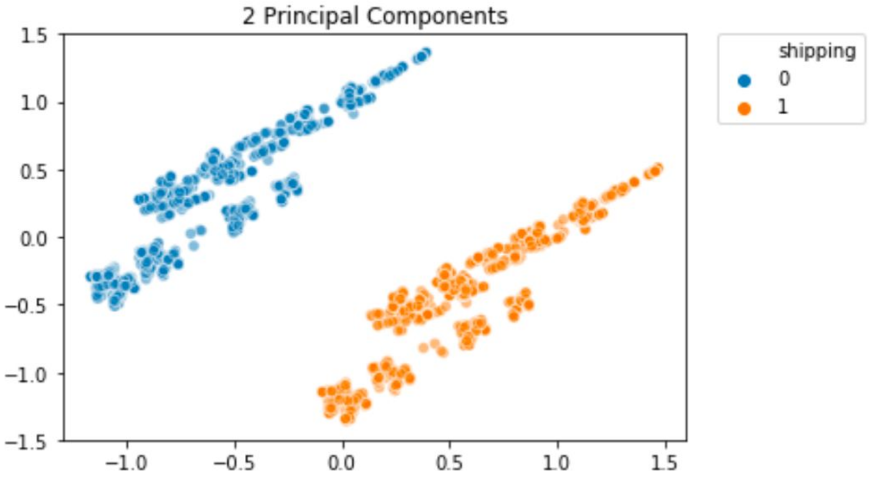


Figure 2 PCA

In this project, we used several models including linear regression, Lasso, random forest, and K Nearest Neighbors (KNN), and then compared model accuracies and time efficiencies among different models.

The inputs for the models are product brands, item conditions, and shipping from the original dataset. The item condition ranges from one to five; one is the worst condition and five is the best condition. And for shipping, zero means that the price listed does not include the shipping fee and one stands for the shipping fee is included. Moreover, we calculated the length of descriptions as a feature for the predictive model. The reason for this is that we noticed items with longer and more detailed descriptions tend to be more expensive compared to other similar products. This could be that sellers who write longer and better descriptions are more motivated to sell the products or the products that are better, (e.g. in better conditions, or have more features) and therefore, need more words to describe. Finally, we used the Vander module in Python to calculate the sentiment scores of the description.

[1] This paper points out that Principal Component Analysis is primarily used for numerical data. http://dx.doi.org/10.4067/S0718-27242013000400008

The Vander module essentially quantifies each item description's sentiment into a score ranging from negative one to one. A negative score means a negative sentiment, scoring a zero mean a completely neutral tone, and a positive score means a positive sentiment. The reason for using sentiment scores is that we believe products that are described as "in *great* condition!" have higher prices that products are described as "okay condition." For other non-numeric data such as names and brands, we used one hot encoding to convert categorical data so that we can use them in the models.

Model Evaluation:

Each model is evaluated based on accuracy and time efficiency. Model accuracy is evaluated based on Root Mean Squared Logarithmic Error (RMSLE), which we abbreviated as MSE. The MSE is defined as

$$\varepsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i+1)-log(a_i+1))^2}$$

Where:
ε is the RMSLE value,
n is the sample size,
p_i is the price prediction for i,
a_i is the actual sale price for i.

III. Implementation

Choosing K for KNN:

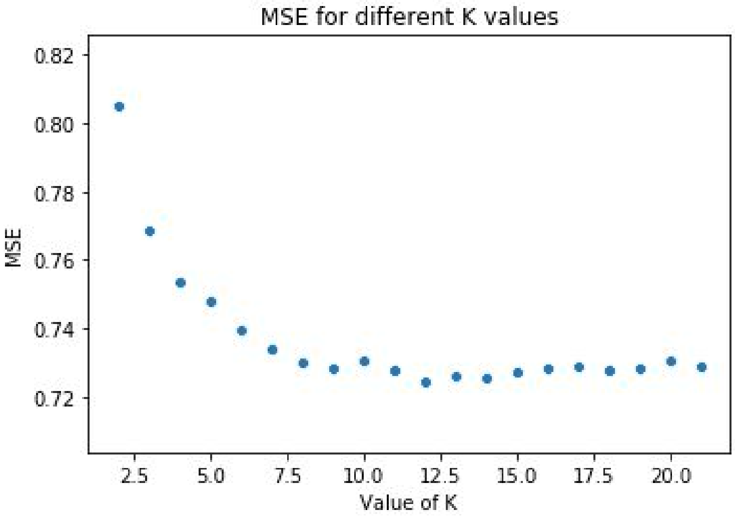


Figure 3 the MSE for different K values

For KNN, the value of K heavily influences model predictions. We wanted to find a K value that yields the lowest error rate. To find the optimal K value, we ran the KNN model on a smaller subset of the data (n = 5,000) with different K values, ranging from 1 to 20 with 0.5 increment. A scatter plot of K values and the corresponding MSE value is shown in Figure 2. When K = 12, MSE is smallest and it is the K value we used.

Time consumption:

Another important factor to consider when it comes to model comparisons is time consumption. Accuracy is essential but we also need to consider the time each model uses. Table 1 lists time consumption for each model for 70,000 (we randomly split data into 70% for training and 30% of data for testing) :

	Linear	**Lasso**	**Random Forest**	**KNN**
Time/s	0.289757	38.885702	63279.48	543.0025

Table 1 Time consumption for each model

From the table above, the Random Forest model uses the longest time and the linear regression uses the least amount of time.

Model Accuracy:

	Linear	**Lasso**	**Random Forest**	**KNN**
MSE	0.725879	0.751482	0.607267	0.656753

Table 2 MSE for each model

From the table above, Random Forest has the best performance for prediction accuracy with MSE at 0.607. Following Random Forest is KNN with MSE at 0.657. Then, lasso and linear regression as accurate as other models. Considering the trade-off between time efficiency and model accuracy, we believe that using KNN with K = 12 would be the preferred method for this problem. Although this model does not do as well as Random Forest, it produces the second best result and is 116 times faster!

IV. Results

Through comparing different models, we found that models rank from linear regression, Lass, KNN (K =12), and Random Forest in terms of time efficiency, with linear regression performed under one second, and Random Forest ran for 17 hours. And models rank from Lasso, linear regression, KNN, and Random Forest from the least accurate to the most accurate. Although Random Forest has the highest accuracy, it takes an excruciating amount of time to run. Considering the time and accuracy trade-off, KNN is the most cost-effective method for this dataset.

It is interesting to note that throughout the project, we found “shipping” is a crucial variable in the dataset. It seems that products can be set apart by whether free shipping is included. Items tend to be more expensive without free shipping. We notice in this dataset, items without free shipping are typically more precious, such as jewelry, smartphones, camera and vintage collectibles. These items could be damaged easily during transportation and require special attention in delivery. Whereas clothes and makeup products are less expensive and are less likely to be damaged during transportation typically include free shipping.