Capstone project 2 : Price Prediction of an Online Marketplace Milestone Report

Contents

Problem statement- why is it useful to answer the question

Online market places have emerged over the course of the last decade. Various marketplaces such as Amazon, Ebay, Etsy, etc. have their own very different business models and how the buyer and the seller interact with each other and also how to set the price of an item in the marketplace. We looked into Mercari marketplace and tried to predict the price of an item based on various features such as item condition, shipping type (seller pays for it or buyer pays for it) and category name.

Clients/ Intended audience

The study is of interest to whoever wants to have an interaction with this marketplace whether it's from the buyer or from the seller side. From the seller side due to the fact that sellers can have the prediction of the price of an item before putting it in the market and from the buyer side, due to the fact that buyers can have the prediction of the price of an item they are looking to purchase.

Dataset used for the investigation

The sets are from www.kaggle.com and the sets are found here: https://www.kaggle.com/saitosean/mercari/version/1

Since the data set is from Kaggle, the test set does not have the price. Therefore, we will only work with the train set and in the Machine Learning phase, we will split the train set into test and train.

The following table summarizes all the variables:

Variable	Description
test_id	the id of the listing
name	the title of the listing as it appears on the item profile
item_condition_id	the condition of the items provided by the sellers; 1 is New With Tags, all the way to 5 as the condition gets worse
category_name	category of the listing (we will split up the category name in this project to create more similarities in seemingly different categories)
brand_name	the name of the brand for the item on sale
shipping	1 if shipping fee is paid by seller and 0 if shipping is paid by buyer
item_description	the full description of the item as it appears on the item profile
price	the price that the item was sold for. This is target variable that we will predict in this study

Data cleaning and wrangling

- Data cleaning: The data is taken from Kaggle. The dataset is pretty clean and the variables are already arranged as columns.

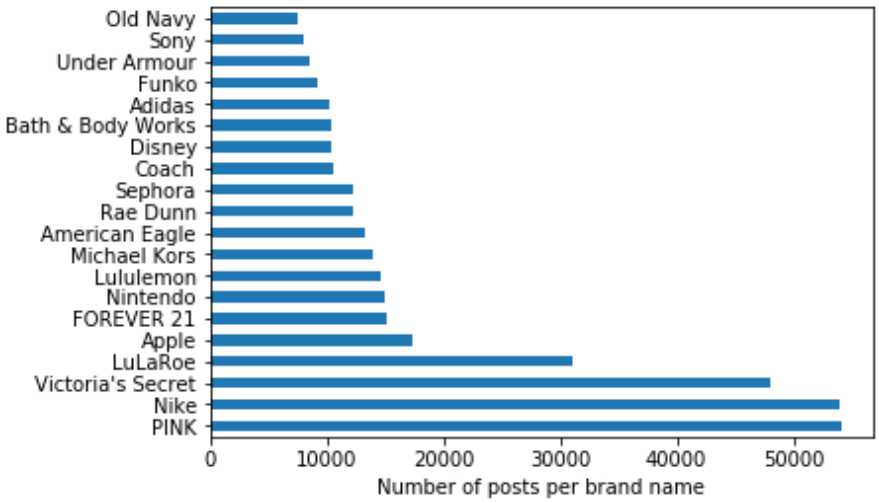
- There are variables with missing values and that there are empty rows for some attributes such as brand name or item category, which we switch to 'no_name'.

  This method to replace all missing values with forward fill and backward fill data.

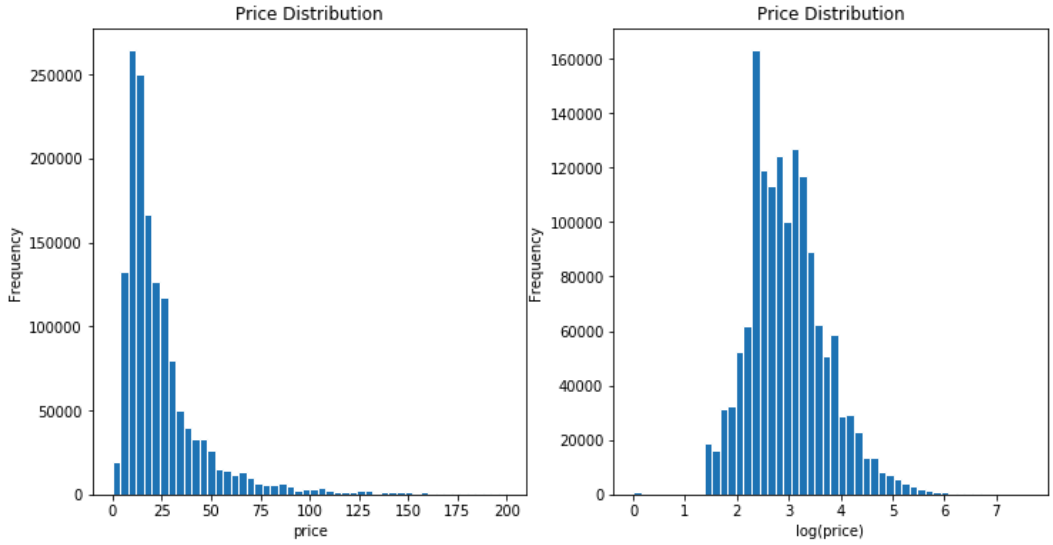
- Outliers: There is a value of $2009 for the price of an item, which we consider an outlier.

Data visualization

Following plot is for the top 20 brand names in the marketplace in the order of the popularity of them in the marketplace.

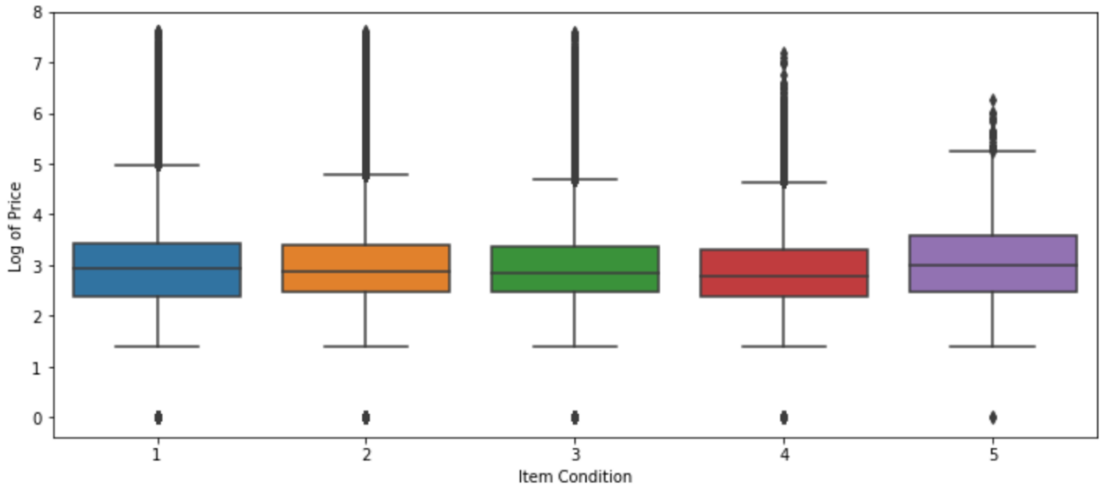


The following figure depicts the distribution of the price and the log of the price in the marketplace.

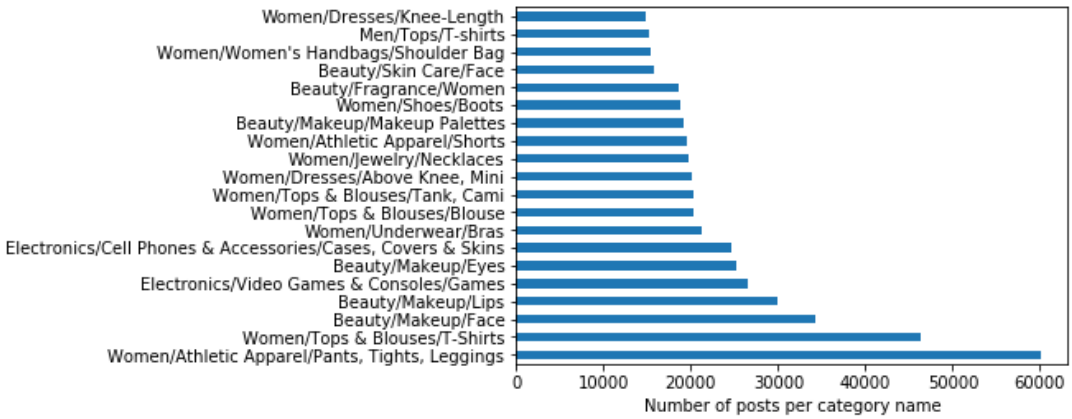


The histogram on the left shows the price is skewed to the right and the histogram on the right shows the log of the price is normally distributed.

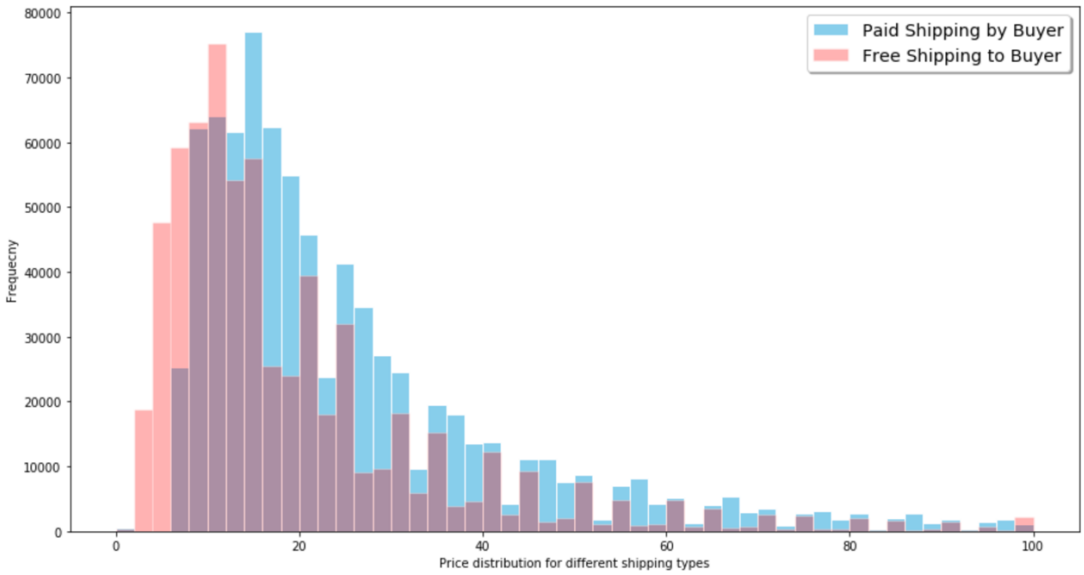
The following box plot shows the distribution of the log of the price and the item condition.



The following is the bar chart of the top 20 category names in the marketplace.



The following is the histogram for the distribution of the price for different types of shipping.



Exploratory Data Analysis (EDA)

Are there variables that are particularly significant in terms of explaining the answer to your project question?

The 'Response' variable is the price of an item that needs to be predicted. As we checked out different price distributions (the box plot of price due to different item conditions as well as histogram of price for different shipping types), there is no significant difference in terms of different variables in the data set having a role in the price prediction.

What are the most appropriate tests to use to analyze these relationships?

- Visual analysis using the box plot, histograms as well as bar charts in order to see the response variable vs the independent variables give us information about the relationships as it is shown above.

Machine Learning Algorithms (ML)

As for the ML part of the project, we convert the variables into dummy variables and values of 0 and 1 and predict the log of the price. In doing so, we first apply Decision Tree regression model and the r^2 score is 0.44. We then switch to Random Forest regression with the accuracy of 85.28%.

Conclusions

In this project, we investigated the odds of predicting the log of the price in the Mercari online marketplace based on different attributes, such as item condition (New With Tags, New, Used, etc.) as well as shipping type (paid by the seller or buyer) and brand name. We used several visualization tools in order to look into our dataset. We saw that the log of the price is normally distributed as opposed to the price itself. We then started the Machine Learning process by using Decision Tree Regressor and then Random ForestRegressor. Between the two models, Random Forest Regressor gave us the accuracy of 85.28%.

The shortcomings in this dataset were that there was very little (almost none) correlation between the `item_condition` and the `price` so the `item_condition` attribute wasn't very effective in our prediction on `price`. Moreover, the `price` itself was rightly skewed so we predicted the log of the `price`.