Predicting Prices of Second-Hand Items on Craigslist with Image and Textual Features

Muhiim Ali, Isabelle Meza, Nuo Wen Lei

Introduction

We are attempting to model prices for second-hand items based on attributes they hold to give sellers an understanding of the expected value of their items.

Hypothesis

We expect there to be a correlation between the price and other attributes of the item.

Dataset

We webscrape from the all-category gallery on Craigslist.

Link to gallery

We collected item image, title, category, price, mileage (for vehicle listings), date, and location.

There are 3,254 item entries with 4 unique days of items between 3/11 and 3/14.

Our dataset contains only publicly available information

Methodology

We preprocess images with a pretrained VGG16 Keras model and then apply PCA to the resulting features to create 210 dimensional image features.

We preprocess item titles with the small text embedding from SpaCy to extract textual features.

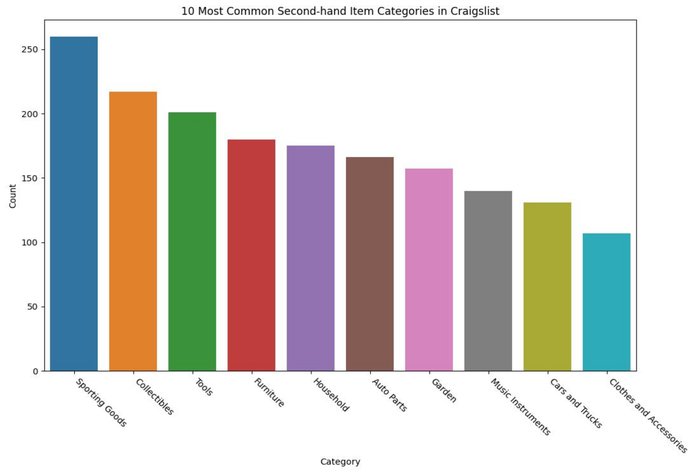
We use the Kruskal-Wallis Test to find if there were any significant differences in mean price between different categories.

We use Linear Regression to see whether the price of items can be linearly predicted from item attributes, which include images, titles, and other details like category and location.

Analysis/Results

1. ***Difference in Mean Price Across Categories:***

We wanted to see if there were any differences in the average prices across different categories. Our null hypothesis was that the average prices would be the same for all categories. To test this, we used the kruskal-wallis test and found a p-value less than 0.05, which means we had to reject our null-hypothesis.



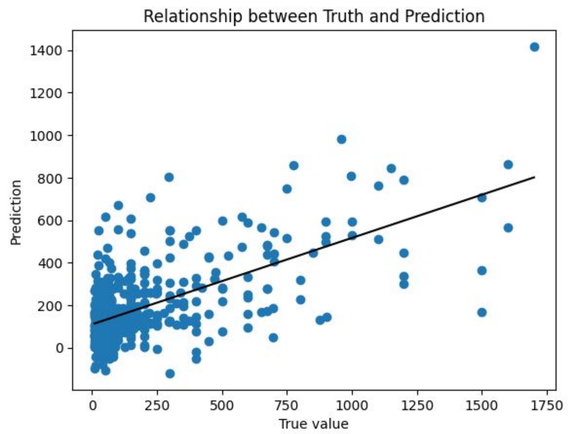
2. Linear Regression Prediction from Item Attributes

In quantitative results, the median RMSE was 262.70, which means that on average the Linear Regression prediction was $262 off from the true price. However the median R2 was slightly positive being 0.08.

The qualitative results display the line of best fit between ground truth and prediction does trend upwards, which signals a positive correlation. This reinforces the quantitative results.

Trial	RMSE	R2
1	256.99	0.1254
2	262.70	0.0383
3	271.00	0.0803
4	600.06	-3.656
5	227.60	0.1585
Median	262.70	0.0802

Table: RMSE and R2 results from Linear Regression over 5 cross validation folds.



Retrospective

Looking back, our hypothesis regarding location and category-based pricing was validated. Although our attempt to predict prices through linear regression gave positive results, it highlights the inherent difficulties in encapsulataing second-hand items' dynamics. Moving forward, our study has the potential to serve as a foundation for further research about the relationship between online resale and the attributes of the items.

We thank Professor De Stefani and the 1951A TA Staff for all their help and support!