Predicting Computer-Associated Item Prices Using Subsetted Linear Regression and Random Forest Regression Models

Jacob Ellison, Aoi Furukawa, Michael Yang
December 13, 2021

UC Berkeley
DATA C100/200: Principles and Techniques of Data Science (Fall 2021)
Graduate Final Project Report

Abstract

Previous research and theories on pricing strategy have focused on deductive approaches, combining relative pricing levels (based on cost, competition, and consumer value), internal objectives (e.g., profit maximization), and external factors (e.g., market conditions). This article explores a more inductive approach, using machine learning models to suggest appropriate pricing based on other products existing in the market. Specifically, it builds multiple linear regression and random forest regression (RF) models that predict prices based on the product metadata and review data from Amazon for computer-related products.

Our result indicates that it may not be feasible to directly model computer part prices since the variability in item type even within the description, and price range, are too high. Despite this, we observed a predictable trend in prices for items with common brands, and items under the $50 price point. Our best performing model was the RF model, which achieved a training and test R-squared value of .925 and .35, and MSEs of $5 and $66, respectively. While clearly this model was overfit, these values support the hypothesis that it is possible to model the price of computer-related items solely from Amazon metadata.

---

1. Introduction

Pricing is one of the most important factors in purchasing decisions by consumers, and hence one of the most important tasks for every business. Previous research and theories on pricing strategy have focused on relative pricing (based on cost, competition, and consumer value), internal objectives (e.g., profit maximization), and external factors (e.g., market conditions) (Kienzler & Kowalkowskia, 2016).

With the growth of e-commerce, there is now more data on products, prices, and consumer reviews than ever before. This data makes possible a more inductive approach, using machine learning models to suggest appropriate pricing based on other products existing in the market. Amazon has adopted this approach in demand forecasting, developing deep learning models to create a unified model for all product categories (Amazon Science, 2020).

This research investigates the following questions: (a) *Can a machine learning model predict the prices of products based on product and review data?* and (b) *What are the most important features for predicting the price?*

2. Dataset Overview

The dataset was obtained from Amazon Review Data (Ni et al., 2019) with more than 233 million reviews for 15 million products in total. A small sample of computer-related products were used, with product metadata for 18,772 items under the 'Computer' category.[1]

The product metadata includes one item per line, with the following metadata for each:

asin	Unique ID for the item	**details**	Additional info about the item
brand	Name of the brand	**feature**	List of features
title	Title of the item	**tech1**	List of item specs
description	Description of the item	**tech2**	List of item specs
price	Price or price range	**also_buy**	List of asin for also bought
rank	Rank of the item	**also_view**	List of asin for also viewed
main_cat	Category of the item	**similar_item**	List of asin for similar items
date	Date of initial listing	**image**	URL to image

The dataset is in JSON format and may have nested values, specifically in `features`, `tech1`, `tech2`, and `descriptions`. A sample of the product metadata is below.

```
{
  "asin": "0000031852",
  "title": "Girls Ballet Tutu Zebra Hot Pink",
  "feature": ["Botiquecutie Trademark exclusive Brand",
              "Hot Pink Layered Zebra Print Tutu",
              "Fits girls up to a size 4T",
              "Hand wash / Line Dry",
```

[1] Dataset retrieved from: https://colab.research.google.com/drive/1Zv6MARGQcrBbLHyjPVVMZVnRWsRnVMpV

```
            "Includes a Botiquecutie TM Exclusive hair flower bow"],
  "description": "This tutu is great for dress up play for your little ballerina.
Botiquecute Trade Mark exclusive brand. Hot Pink Zebra print tutu.",
  "price": 3.17,
  "imageURL": "http://ecx.images-amazon.com/images/I/51fAmVkTbyL._SY300_.jpg",
  "imageURLHighRes": "http://ecx.images-amazon.com/images/I/51fAmVkTbyL.jpg",
  "also_buy": ["B00JHONN1S", "B002BZX8Z6", "B00D2K1M3O", "0000031909",
"B00613WDTQ", "B00D0WDS9A"],
  "also_viewed": ["B002BZX8Z6", "B00JHONN1S", "B008F0SU0Y", "B00D23MC6W",
"B00AFDOPDA", "B00E1YRI4C", "B002GZGI4E", "B003AVKOP2", "B00D9C1WBM"],
  "salesRank": {"Toys & Games": 211836},
  "brand": "Coxlures",
  "categories": [["Sports & Outdoors", "Other Sports", "Dance"]]
}
```


3. Data Processing

The dataset had numerous missing and extraneous values that needed to be removed.

3.1. Missing Price

Since the price data is required to build a price prediction model, items without price data were removed. 2 items with a range (e.g., $10.00 ~ $15.00) were also removed. The resulting dataset had 2,648 items.

3.2. Missing Rank

From the previous dataset, 117 items (~5%) did not have ranking data. These items were removed, resulting in 2,531 items.

3.3. Extraneous Title

Of the remaining dataset, 16 items had titles with extraneous HTML/CSS code instead of titles. These items were removed, resulting in 2,514 items.

The processed dataset had 2,514 items and 16 columns. 100% of the items had `title`, `price`, `rank`; 95% had `brand` and `date`; 63% had `description`.

Figure 1. *Proportion of Missing Values*

Columns	Missing Values	% of Total
description	13642	73%
title	1	0%
image	4688	25%
brand	477	3%
rank	692	4%
main_cat	0	0%
date	184	1%
asin	0	0%
feature	14665	78%
tech1	15186	81%
also_buy	17337	92%
price	16122	86%
also_view	16472	88%
tech2	17113	91%
details	18746	100%
similar_item	18409	98%

4. Feature Engineering

The following columns were featurized:

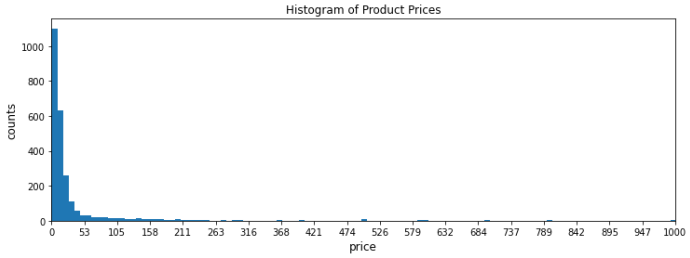
- `price` as metric dependent variable
- `rank` and `date` as metric variables
- `brand` and `subcategory` as categorical variable
- count of keywords in `title` and `description` as a metric variables

4.1. Price

Item prices had a heavily skewed distribution, with a maximum value of $9,018 and median value of $10. The 99th percentile price was approximately $700 and $3,380 for 99.9th percentile.

3 items with outlier price above $3,000 were removed.

Figure 2. *Distribution of Item Prices*



4.2. Rank

Ranking is an important metric for the item's popularity, which in turn is a proxy of how 'acceptable' the price is for the item.

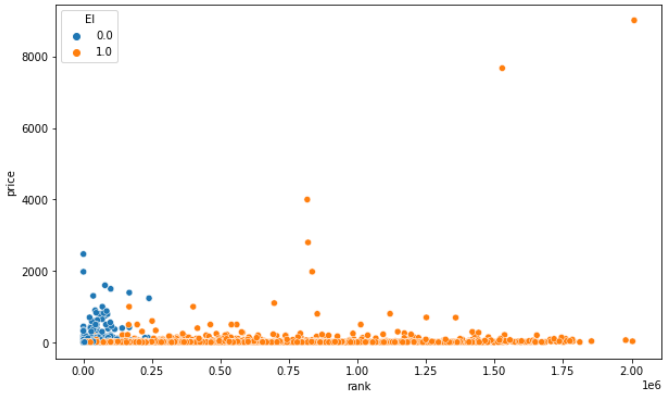
`rank` was extracted from the string format and converted into a float as a metric variable.

4.3. Subcategory

The category for all items in main_cat was 'Computer'. However, while extracting the rank data, it was found that some rankings were under 'Electronics'.

EDA showed a potentially meaningful clustering, which led to featurization of this data as a categorical variable of 1 for 'Electronics' and 0 for 'Computers & Accessories'.

Figure 3. *Price and Rank of 'Computer' and 'Electronics'*



4.4. Date

Date of listing signifies how long the product has remained listed. It is possible that the 'age' of product listings are negatively correlated with price because it has become outdated, or they are positively correlated because it is a steady seller. In both cases, it may have a meaningful association with price.

`date` was converted into integer of days as a metric variable, by subtracting the date of the oldest item. The variable signifies the relative 'age' of the item.

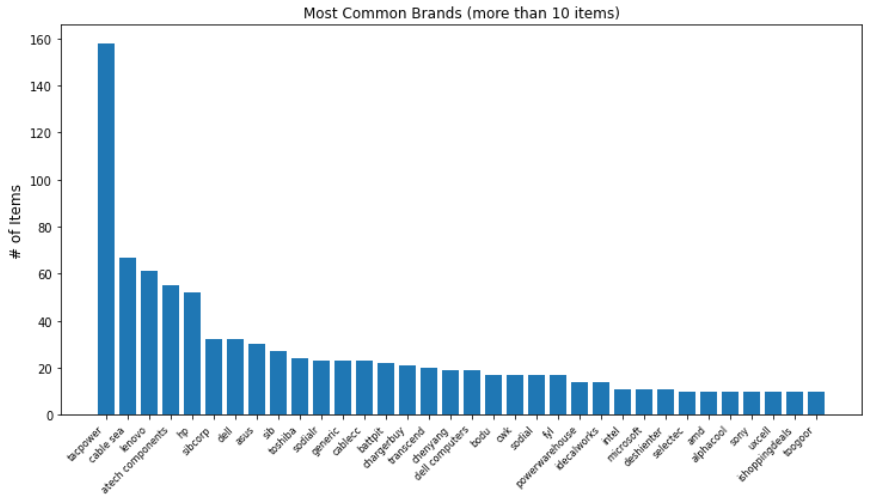
4.5. Brand

Brand is an important factor of any product. This was quantitatively confirmed in a similar machine learning model for suggesting prices of items on the e-commerce platform Mercari (Kumar, 2020). Premium brands can be a proxy to qualities such as 'reliability' and command higher prices, whereas unknown brands do not have such trust from consumers and compete by offering the lowest possible price.

The dataset had 1,117 brands, of which 73% (793) had only 1 item listed, and 98% (1025) had less than 10 products listed. Brands with less than 10 products have none or little brand awareness among consumers, while adding noise to the model. These brands were encoded as `No Brand`.

34 brands had more than 10 items listed. Some were lesser known yet prolific brands like Tacpower and Cable Sea that primarily sell computer accessories, and some were well-known brands like Lenovo, HP, Dell, and Asus.

Figure 4. *Brands with the Most Number of Items*



4.6. Keywords

Words are the only way for the machine learning model to 'understand' what the item actually is. Finding the relationship between words and price is a core part of building this model.

As seen in Section 4.3, 100% has title and 63% has description within the filtered dataset. These values were parsed and tabulated into unique word counts to identify potential keywords.

Of more than 26,000 unique words, 21 keywords were chosen (17 for `title` and 14 for `description`) as our set to featurize. The keywords for selection are as follows:

- *Laptop, Memory, RAM, Chip, Fast* signify computers
- *Touchscreen, Touch, Controller, Headset, Mouse, Keyboard* signify electronics
- *Charger, Case, Adapter, Power, Card, Software* signify accessories
- *Home, Business, School, Backpack* signify context of use

`feature` (52%), `tech1` (25%), and `tech2` (9%) also have data but have many more missing values. We investigated these features but opted not to featurize them as it had too many missing values.

5. Methodology

5.1. Initial Evaluation with Multiple Linear Regression Model

Multiple linear regression model was chosen for an initial evaluation of our features, as the model needs to predict a continuous metric value of price.

Due to the wide variety of items and price became evident during the EDA, we knew that the linear regression model may have low accuracy and may need subsetting of the items to improve its prediction accuracy.

5.1.1. Subsetting Strategies

Since the spread of price was heavily skewed, we investigated subsetting by price. The maximum below outliers at 75th percentile price was $50, so we tried to subset the data by below or above $50. Since the median price was $10, this subsetting by dependent variable did not significantly improve the model.

5.1.2. Keyword Optimization Strategies

We used inter-item correlation to check for any keywords that may be correlated with each other, confounding the impact of each word. The visualization did not show clear patterns, confirming that the words chosen did not show high correlation.

We also plotted the correlation coefficients of each keyword and brand categorical variables to visualize and identify the most significant terms. This approach can be replicated in subsequent iterations of keyword selection and optimization.

Figure 5. *Inter-feature Correlation Heatmap*

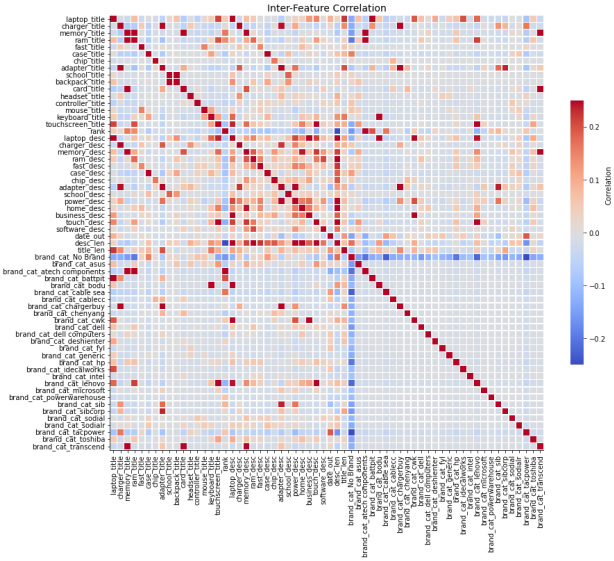
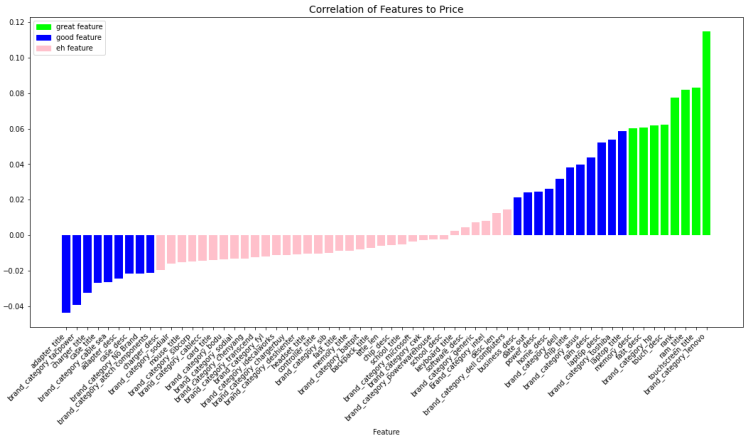


Figure 6. Coefficients of Features



5.2. Subsetting Optimization with PCA and Decision Trees

Given the wide variety of items in this dataset for the 'Computer' category, PCA was conducted without price data, in order to investigate whether there were natural clusters among the items.

The visualization of the first 3 principal components showed two clear clusters in our dataset.

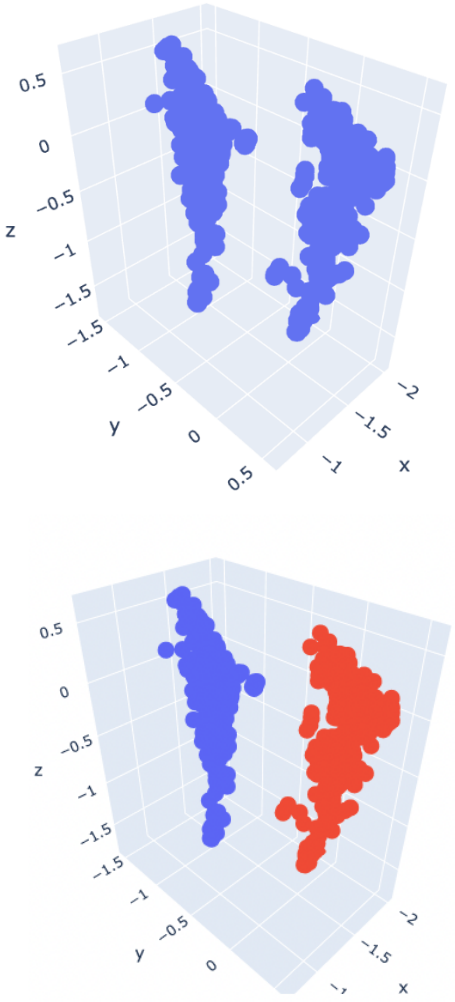


Figure 7. *PCA Plot, before and after subsetting (Right)*

The feature that led to such clustering could significantly improve the model. In order to identify this feature, we created a decision tree with depth of 1 and iterated over each feature and calculated entropy.

The single feature with maximum entropy which led to a perfect split of the clusters was whether the items had a brand, instead of "No Brand" as featurized above as brands with less than 10 items listed.

Subsequently, we subsetted the data by "Common Brand" and "No Brand", and developed 2 models for each cluster.

5.3. Random Forest Regression Model

The random forest regression model (RF model) performs well on regression tasks, especially with datasets with nonlinear patterns and when the input data contains many discrete random variables that can form natural decision criteria for trees.

We chose to deploy RF model for our price prediction model, because our dataset has a variety of categorical variables encoding brands, which was revealed to be a significant predictor in our PCA.

In order to test this we developed a model on our final data, tiered by the inclusion criteria of price and having a well known brand and tested a RF model.

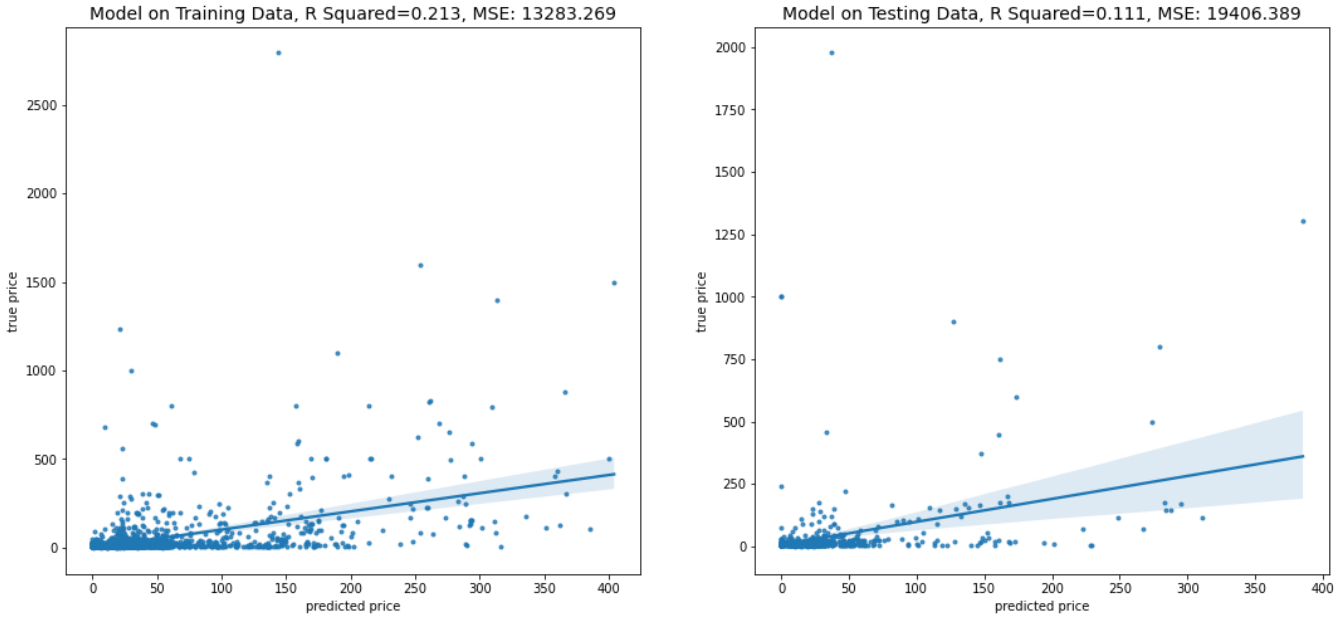
Hyper-parameter optimization was conducted through a grid search with 3-fold cross validation, through the max depth [10, 20, 30, 40, 50, 100] and number of estimators [30, 50, 100, 200]. This resulted in the optimal hyper-parameter as 200 max depth and 100 estimators.

6. Results

6.1. Initial Multiple Linear Regression Model

The initial multiple linear regression model (initial model) had the lowest accuracy, with This was expected, as the variability of price and types of items in the dataset became clear during the EDA.

Figure 8. *Performance of Initial Model*

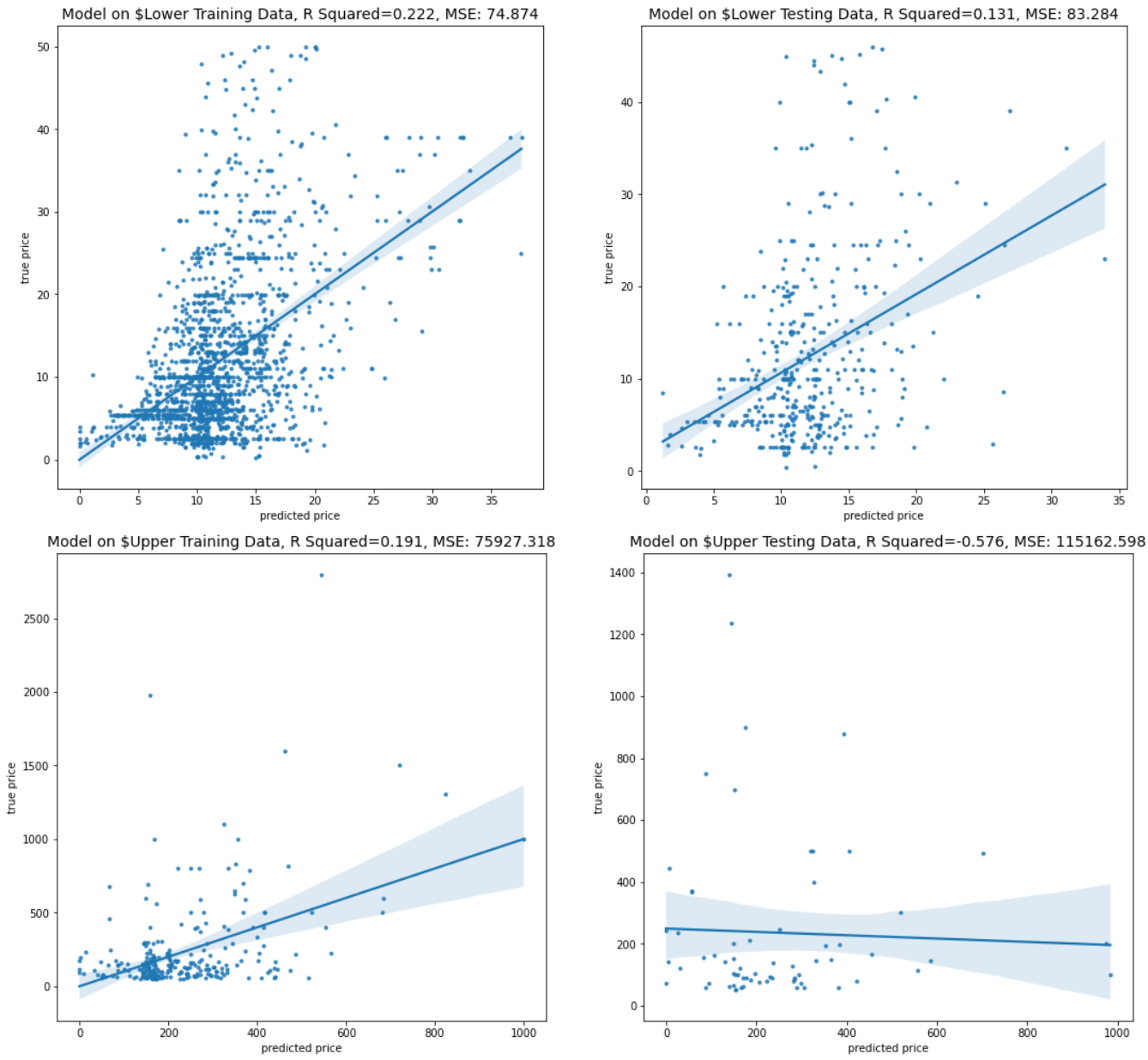


(Continued)

6.2. Multiple Linear Regression Model with Price Subsetting

Subsequently, we subsetted the data by price threshold of $50 on the initial model. While there was a small increase in accuracy, it was not significant.

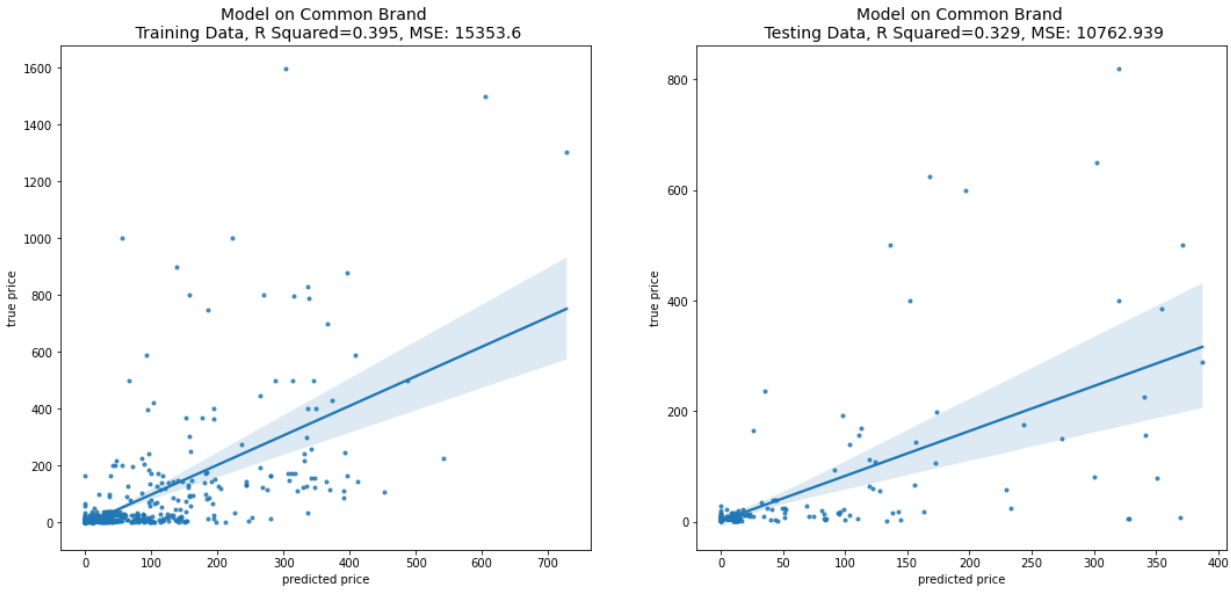
Figure 9. *Performance of Initial Model with Price Subsetting*
(Under $50 is $Lower, over $50 is $Upper)



6.3. Multiple Linear Regression Model with Brand Subsetting

Through PCA and subsequent analysis using decision tree, we identified that "Brand" was a significant categorical variable. Subsetting the data by this metric (brands with more than 10 items as "Common Brand" and less as "No Brand") and applying our initial model showed a significant improvement in accuracy. This indicated that the approach of using this feature as a criteria for modeling was necessary.

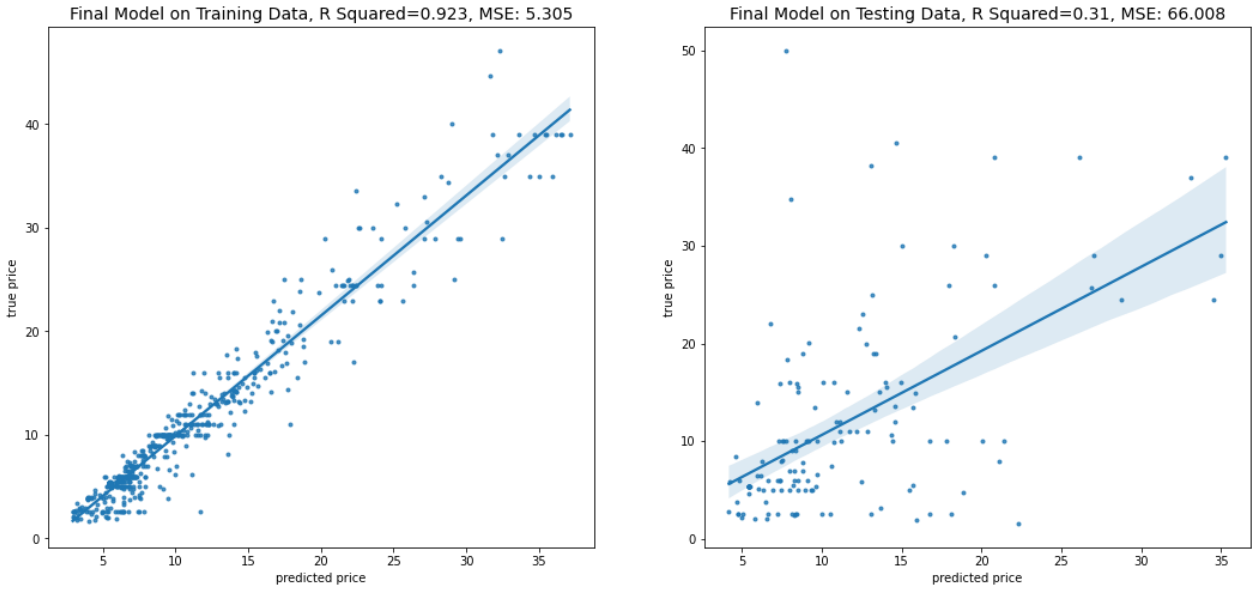
Figure 10. *Performance of Initial Model with Brand Subsetting*
(Less than 10 items as "No Brand" and more than 10 items as "Common Brand"



6.4. Random Forest Regression Model with Hyperparameter Optimization

With the understanding that brand is a significant categorical variable, we subsetted the data by brand and price thresholds from previous iterations of the model.

Figure 11. *Performance of Final RF Model with Brand and Price Subsetting*



7. Discussion

7.1. Limitations

Even though the model has improved through implementation, there might be an overfitting problem. Since the dataset we use is not large (train set 565, test set 142) and has a lot of missing values, the model might not work well for larger data.

While we used train_test_split to avoid overfitting in this study, the improvement had a limitation because of the small data size. To overcome this problem, we would like to find other datasets to train and validate the model for future study.

Also, the dataset "Computer" contains a variety of products, which makes it difficult to predict prices. We expected that there would be a lot of computers in and we could predict prices with its functions extracted from its description. However, most of the dataset is about cheap accessories and it is difficult to predict price as there are not so many differences between functions of accessories.

As for the impact of the study and ethical concern, while the model we created would help sellers to determine their products' price, if the price prediction model like this one were to be widely used in society, there might be some people who would purchase products for speculative purposes and might cause shortages of specific products.

7.2. Future Work

As we mentioned in the previous section, the size of the dataset is one thing we would like to improve. We could improve the current model with a larger dataset. Also, if there are fewer missing values, we could add other features as well. For example, since the original dataset was collected to analyze reviews, we could combine review data and price data and see if there are different trends for word frequency in lower price products and upper price products. Another thing we could explore in the future would be time series analysis. Since the data has date values, we could follow the changes in prices.

8. Conclusion

From our EDA and modeling, we observed that while infeasible to model the entire range of products in the computer parts category, we were able to produce a model that performed with an R-squared value of .34 on our test set, when subsetting the data to include items below $50 and items with well known brands. Our final MSE on the test set was $66.

Our result indicates that it may not be feasible to directly model computer part prices since the variability in item type even within the description, and price range, are too high. Despite this, we observed a predictable trend in prices for items with common brands, and items under the $50 price point. Our best performing model was the RF model, which achieved a training and test R-squared value of .925 and .35, and MSEs of $5 and $65, respectively.

This results indicate an overfitting issue, which means variance of model is high and bias is low. We address this with ensembling to lower the validation error and variance. Nonetheless, these values support the hypothesis that it is possible to model the price of computer-related items solely from Amazon metadata. With a bigger dataset and better features, the approach explored in this project can be iterated to create a price prediction model.

References

Jianmo Ni, Jiacheng Li, Julian McAuley (2019). *Justifying recommendations using distantly-labeled reviews and fined-grained aspects*. Empirical Methods in Natural Language Processing (EMNLP). Retrieved from https://cseweb.ucsd.edu/~jmcauley/pdfs/emnlp19a.pdf.

Amazon Science (2020). *The history of Amazon's forecasting algorithm*.

S Won Lee (2020). *Predict the price of products using Machine Learning*. Medium.

Andreas Hinterhuber, Stephan Liozu (2012). *Is It Time to Rethink Your Pricing Strategy?* MIT Sloan Management Review.

Mario Kienzlera, Christian Kowalkowski (2016). *Pricing strategy: A review of 22 years of marketing research*. Journal of Business Research. https://doi.org/10.1016/j.jbusres.2017.05.005

https://towardsdatascience.com/mercari-price-suggestion-97ff15840dbd

Arun Kumar (2020). *Price Prediction using Machine Learning Regression — a case study in Mercari Price Prediciton Challenge*. Towards Data Science.