Depop Seller Product Pricing

Objective

This project aims to delve into analysis and predictive models in order to help sellers on the popular clothing resale app "Depop" be able to accurately price their items. Predictive models can offer a way to gain competitive, optimized pricing for sellers, especially for those new to the platform or unfamiliar with what might currently be trending. Factors that are most important to consider when pricing an item include trendiness, brands, quality, and seasonality. All of these can greatly influence the market value of a product, making them important to consider when listing an item. Having competitive pricing strategies ultimately increases the chances of having more successful sales, thus driving seller profit. Using data-driven prediction techniques allows for time and effort saved for all sellers, whether you're new to the platform or experienced. Guidance when trying to determine a price point can especially be helpful for sellers with a diverse range of items, as less common items can be trickier to price. Overall, the objective of this project is to empower sellers who may not be as familiar with the market, and guide them to be able to effectively price their items based on the product's attributes, ultimately leading them to find success in reselling.

Data Key, Cleaning, & Encoding

- ➔ Datetime: Date and time the product was posted
  - ◆ No cleaning
- ➔ Link: Unique link to the product
  - ◆ No cleaning, used as a unique identifier to remove duplicates
- ➔ Price: Current, full price the item is listed for
  - ◆ Removed '$' symbol making the variable fully numeric, removed outliers using z scores
- ➔ Description: Description of the product provided by seller on listing
  - ◆ No cleaning
- ➔ Brand: Brand of the product
  - ◆ If brand was "NULL", replaced with "None"
- ➔ Condition: Condition of which the product is in
  - ◆ No cleaning, removed values where condition was "NULL"
- ➔ Title: Title of the product as listed by seller

- ◆ No cleaning, extracted last word and made it a new categorical column called “Type”

➔ Size: Size of the item
- ◆ No cleaning, if sizes were listed as “Multiple Sizes”, it meant the seller was most likely dropshipping (buying multiple products new from a cheaper website and reselling them for more), so binary column was created called “Dropshipping”

➔ Amt.Sold: Amount of items the seller has sold in total
- ◆ Removed rows that did not contain the word “Sold” (formatted wrong), then extracted the amount sold and made the value numeric

➔ Activity: When the seller was most recently active on the app
- ◆ No cleaning

➔ Recent.Review1: The seller’s most recent customer review
- ◆ No cleaning

➔ Time.Listed: How long ago the product was listed
- ◆ Formatted to numeric

➔ Sold.and.Reviews: How many shop reviews the seller has and what their total average rating is out of 5 stars
- ◆ Converted to a numeric column, which provides the seller’s rating out of 5 stars

➔ Discount: If the product offers a discount, what % off it is
- ◆ Extracted the numeric discount, 0 if discount was “NULL” making the column numeric

➔ In.Bags: How many people’s bags the item is in currently
- ◆ Extracted the number of how many bags the item was currently in, making the variable numeric

➔ Like: How many likes the item has currently (up to 99)
- ◆ If Like was “NULL” change it to 0, making it fully numeric

➔ Free.Shipping: If the item has free domestic shipping or not
- ◆ Changed to binary, 1 for free shipping, 0 for no free shipping

➔ Color: Color of the item

➔ Material: Material of the item/Style

➔ Style: Style of the item
- All three style variables (Color, Material, and Style) were changed to binary and made into separate columns for each unique value they contained.

Data Engineering

Changing the format and engineering new variables was key in the analysis of this dataset. A lot of variables needed to be extracted and converted to numeric in order to be included in the models and analysis. Including dummy variables also proved to be important in the process of building models, as variables like brand and condition contributed greatly to overall performance.

Challenges

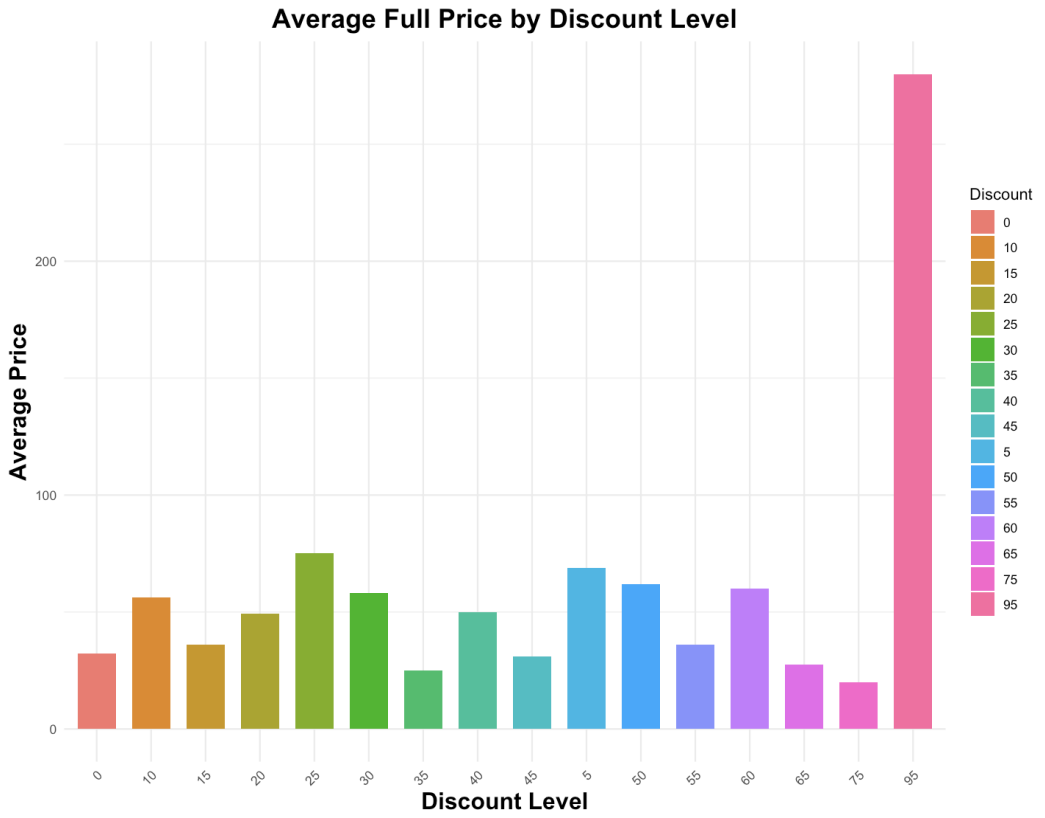
Some troubles that came about when attempting to transform variables included the use of sentiment analysis. Sentiment analysis was ran on the description and recent review variables, hoping to be able to assess whether each was positive or negative. However, it actually increased the root mean square error on all of the models, so it was not included or used in any analysis or predictions.

Another challenge faced was the decision of whether or not to include the 'Like' variable in the analysis and prediction. As a 'like' is always given after a product is posted, it wouldn't be valuable in predicting the price of a new product. However, since Depop contains profiles and has a social media-like platform, with followers and branding, users can average their likes across previous posts to determine how much exposure their new product is likely to get. Users with more followers will have more likes, and if likes are shown to contribute to pricing strategies, this can help the seller determine pricing based on things they already know, such as follower and like count.

Exploratory Analysis

Free Shipping

Seller's often use the offer of "free shipping" to create an incentive for buyers, and also to create the illusion of a cheaper price. This pricing strategy can be extremely effective, and allow for more profit. Adjusting the base price to compensate for the free shipping can not only drive sales, but maintain the overall cost. Shown in the graph above, sellers on Depop are using this strategy as the average price for items with free shipping is significantly higher than those without.



The same technique is also used for discounting products. Sellers are able to leverage profits by pricing items higher, but then offering a discount to make the customer feel like they're getting a good deal when in reality, they're paying what would be full price. This dynamic pricing strategy gives sellers a leg-up in marketing.

Preliminary Model

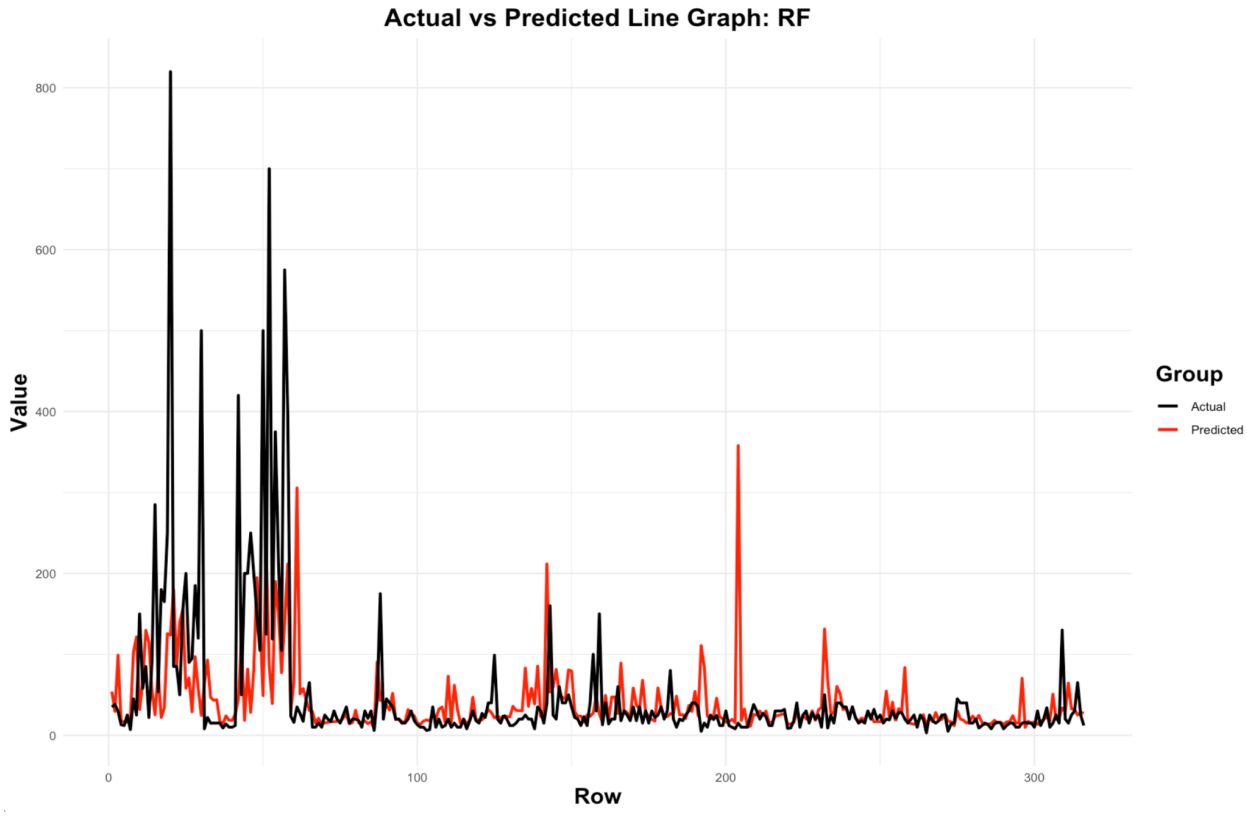
Linear regression was applied on the clean data to determine which variables were the most significant and their predictive capability on the price . This model yielded a multiple R-squared of 78%, which shows the variables demonstrated a strong predictive power. The variables with the lowest P-values included how many likes the product got,

how many people's bag the product is in, and the discounts a product has. This indicates that these are the most significant variables when trying to determine the price point, and should be looked at carefully.

Predictive Model

Random forest and ridge regression were chosen models due to the fact that the price variable is continuous and numeric. The random tree model was found to have the most accurate predictions when used on the cleaned data set, yielding an root mean squared error of around $33. While this number isn't ideal compared to the overall average price of $34, most of this error comes from high-priced items, which all models were shown to be sub-par at predicting.

The random forest model is found to be most accurate when predicting lower values, rather than higher values. This can be attributed to the range of prices in the data not being evenly distributed. It is assumed that if more data was collected, the root mean squared error would go down significantly and improve model performance.



As seen in the graph above, which depicts the actual price compared to predicted, high prices are either falsely predicted, or often missed. Also shown is the fact that the prices aren't evenly distributed, as higher prices appear more often towards the beginning. However, both lines follow a loosely similar pattern, and lower data points are shown to have far less discrepancies compared to higher priced items. Given more data the model would perform better, even on items that exhibited abnormal pricing patterns.

In the context of predicting pricing, a one to four dollar pricing error isn't going to make a significant difference in practicality and ability to sell. This is a reason why this model is best suited for non-luxury items, and items looking to be sold below a fifty dollar price point.

Future Analysis

In order to further analyze, gathering more data would be ideal in order to be able to further explore and enhance model performance. As well as this, being able to predict other variables would be helpful, such as like count. For a less precise prediction model, transforming the price variable into categorical buckets (e.g. $10-$20, $20-$30), could be beneficial and offer sellers a different perspective on pricing, and also has the chance to provide better accuracy.

Analyzing the time of year that the item was posted and comparing it to the type could also give insight into how seasonality plays a role in pricing. Asking if sellers are pricing jackets higher in the winter, or tank tops higher in the summer, can give sellers insight on when to post a product or when to offer a discount.