Women's Blouses Resale Value Modeling

Caitlin Beale
Department of Atmospheric and Oceanic Science, University of California Los Angeles

AOS C111: Introduction to Machine Learning for Physical Sciences

Dr. Alexander Lozinski

Introduction

The goal of this project is to explore and predict the resale value of clothing items using machine learning techniques, leveraging the Mercari Price Suggestion Challenge dataset. Specifically, this analysis focuses on the category "Women/Tops & Blouses/Blouse," with the aim of assisting environmentally-conscious consumers and sellers in making informed decisions about which clothing items retain high resale value. The dataset includes various features, such as item descriptions, prices, shipping costs, and brand names. Linear and Ridge Regression analysis, and Decision Tree models will be employed to complete this task. These models will be evaluated by their R^2 value to determine their ability to predict the optimal resale prices of blouses.

Data and Preprocessing

This dataset contains a multitude of product categories such as:

- 'Men/Tops/T-shirts'
- 'Women/Tops & Blouses/Blouse'
- 'Handmade/Jewelry/Clothing'
- 'Vintage & Collectibles/Supplies/Ephemera'

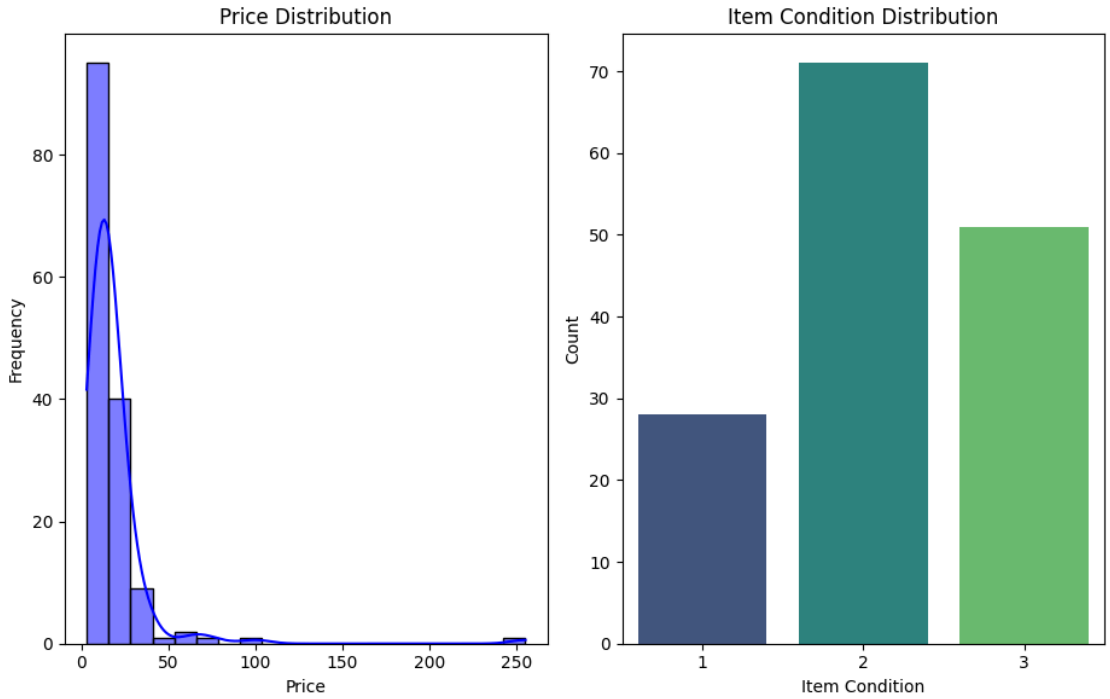
Ultimately, I decided to narrow down the data to analyze just the 'Women/Tops & Blouses/Blouse' category, which left 15,026 entries. Next, I randomly sampled 150 data entries in order to make processing faster. The table below illustrates the first look statistics about these entries.

Table 1:

	Item Condition ID	Price
Mean	2.15	17.74
Minimum	1	3
Maximum	3	255
25%	2	10
50%	2	13
75%	3	18

The charts below (Figure 1) illustrate what the table says above. Notably, most of the resale prices fall around $17. Item conditions are valued between 1 and 3, with 1 being the best condition and 3 the worst. Most items were listed with a condition of 2.

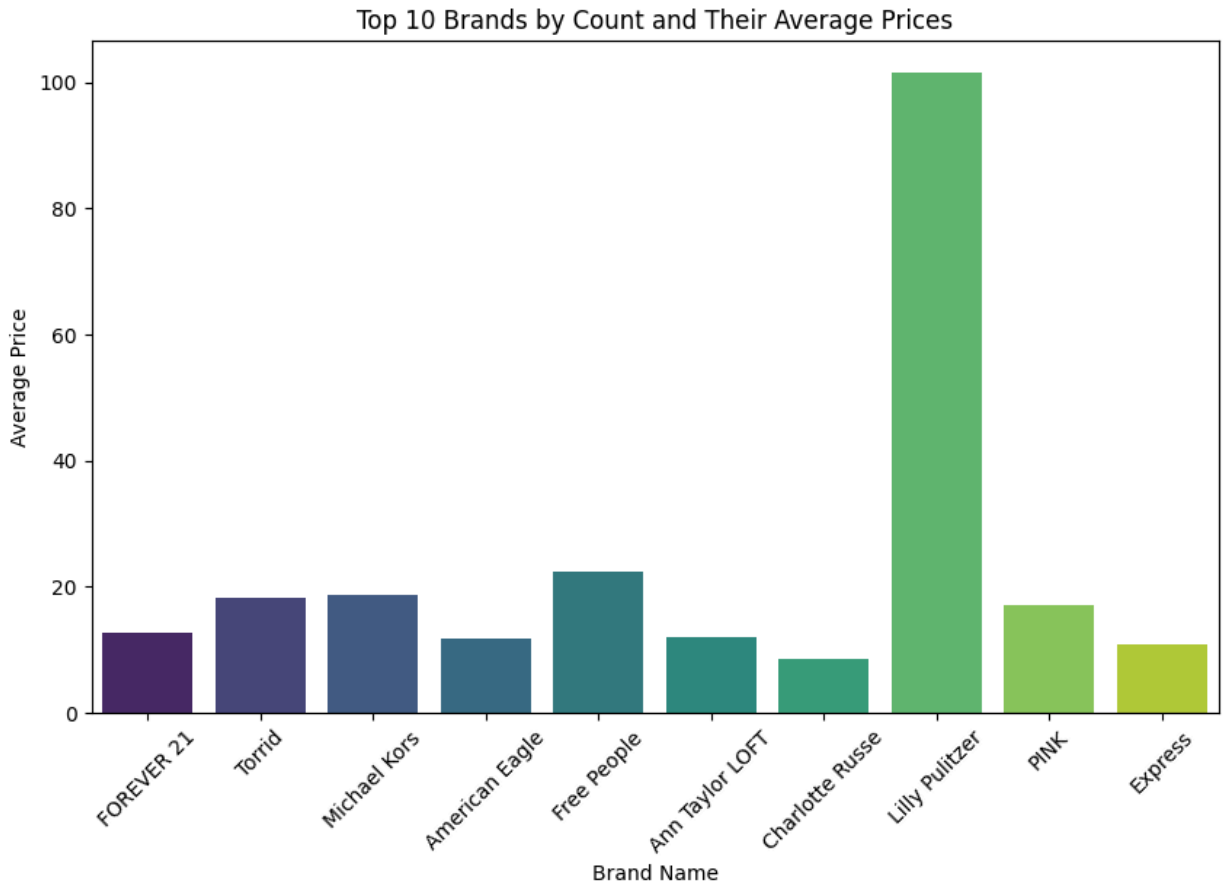
Figure 1:



Categorical features, such as brand name and item description, posed challenges for model training because machine learning algorithms require numerical inputs. To address this, one-hot encoding was used for the initial model. One-hot encoding converts categorical variables into a set of binary variables. For example, each unique brand in the dataset becomes its own column, and the value is set to 1 if the row belongs to that brand and 0 otherwise. This approach was chosen because it allows the model to handle non-numerical data without imposing an arbitrary ranking system that could bias the results. However, one-hot encoding has limitations, particularly when the dataset contains a large number of unique values, as was the case with brand names in this dataset. This method significantly increases the dimensionality of the data, making it computationally expensive and potentially leading to overfitting, especially with smaller datasets.

The chart below (Figure 2) displays the top 10 most frequent brands within the 150 entry sample and the products' average price for that brand. It should be noted again that most prices fall around the $17 mark with the exception of Lilly Pulitzer which has an average price of almost $100.

Figure 2



Modeling

Linear Regression was tested first as a baseline model to track improvements made by more sophisticated techniques. This choice was intentional, as Linear Regression is straightforward to implement and serves as a benchmark to evaluate the impact of modifications and alternative algorithms. The initial results provided a reference point for understanding how well the dataset could be modeled with minimal preprocessing and parameter tuning.

Ridge Regression and Decision Trees were chosen for this analysis because they address distinct challenges presented by the dataset. Ridge Regression was selected because of its ability to prevent overfitting through regularization. By penalizing large coefficients in the model, the Ridge approach reduces the risk of overfitting, particularly in datasets with many features, such as those created by one-hot encoding. This is especially important for this analysis because the dataset includes a mix of numerical and categorical features, some of which are sparse. To optimize the Ridge Regression model, the alpha parameter was tuned using RandomizedSearchCV. The alpha parameter controls the strength of regularization, balancing the trade-off between model complexity and fit. RandomizedSearchCV was used because it is computationally efficient for finding the optimal alpha compared to an exhaustive grid search.

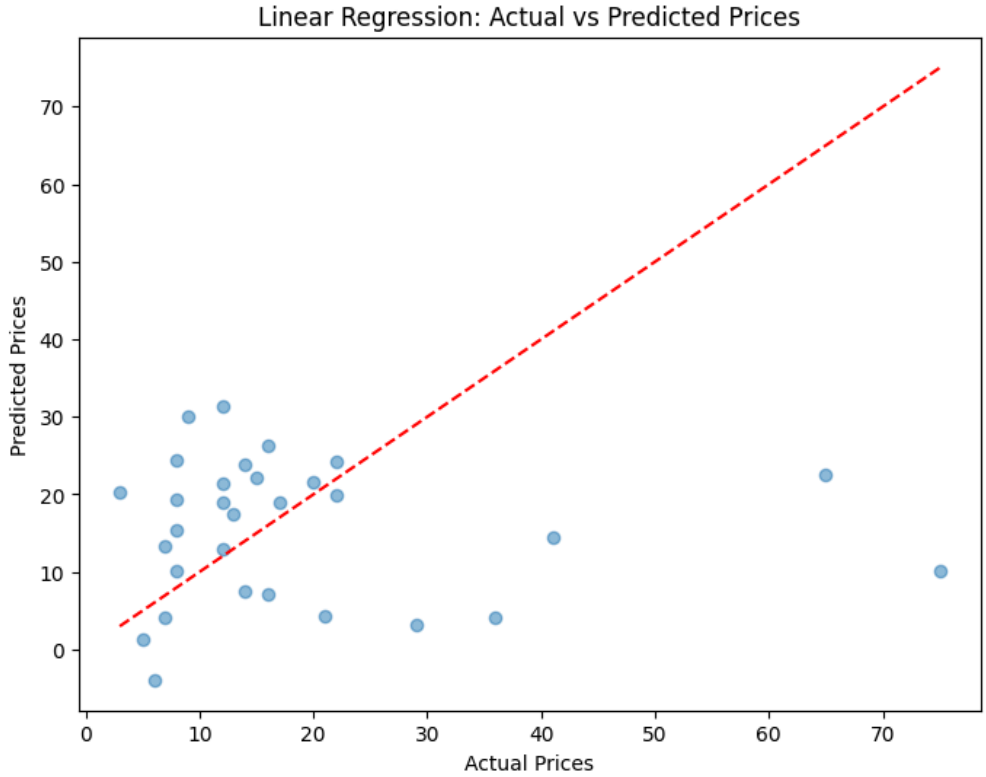
Decision Trees were chosen for their ability to model non-linear relationships and capture complex patterns in the data. They excel at identifying interactions between features, making them well-suited for analyzing the relationships between brand, condition, and price. However, decision trees are prone to overfitting, especially on small datasets, so parameter tuning was essential.

To validate the models, the dataset was split into 80% training data and 20% testing data. This split ensures that the models are trained on a majority of the data while still leaving a significant portion for evaluation, providing an accurate measure of how well the models generalize to unseen data.

Model Results

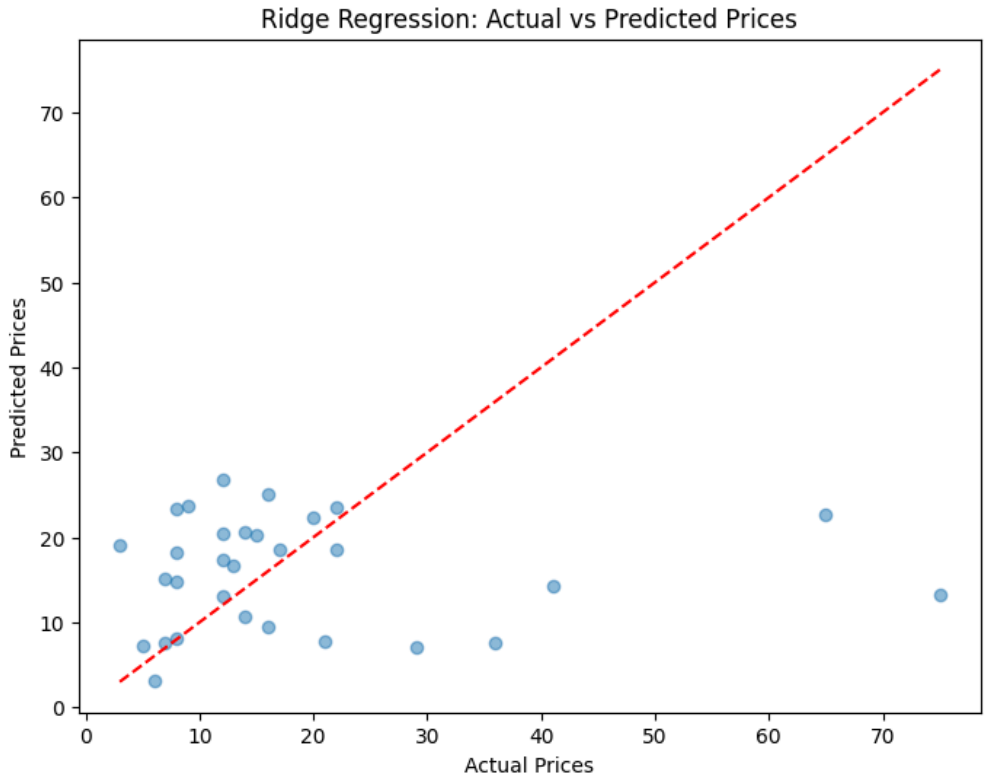
To evaluate the accuracy of each model tested I utilized an R^2 score. For R^2 scores, the closer the score is the one, the more accurate the model is performing. Without adjusting any parameters, the Linear Regression presented an R^2 score of -0.3786, not a very desirable score. The results of the Linear Regression model can be seen in Figure 3 below.

Figure 3:



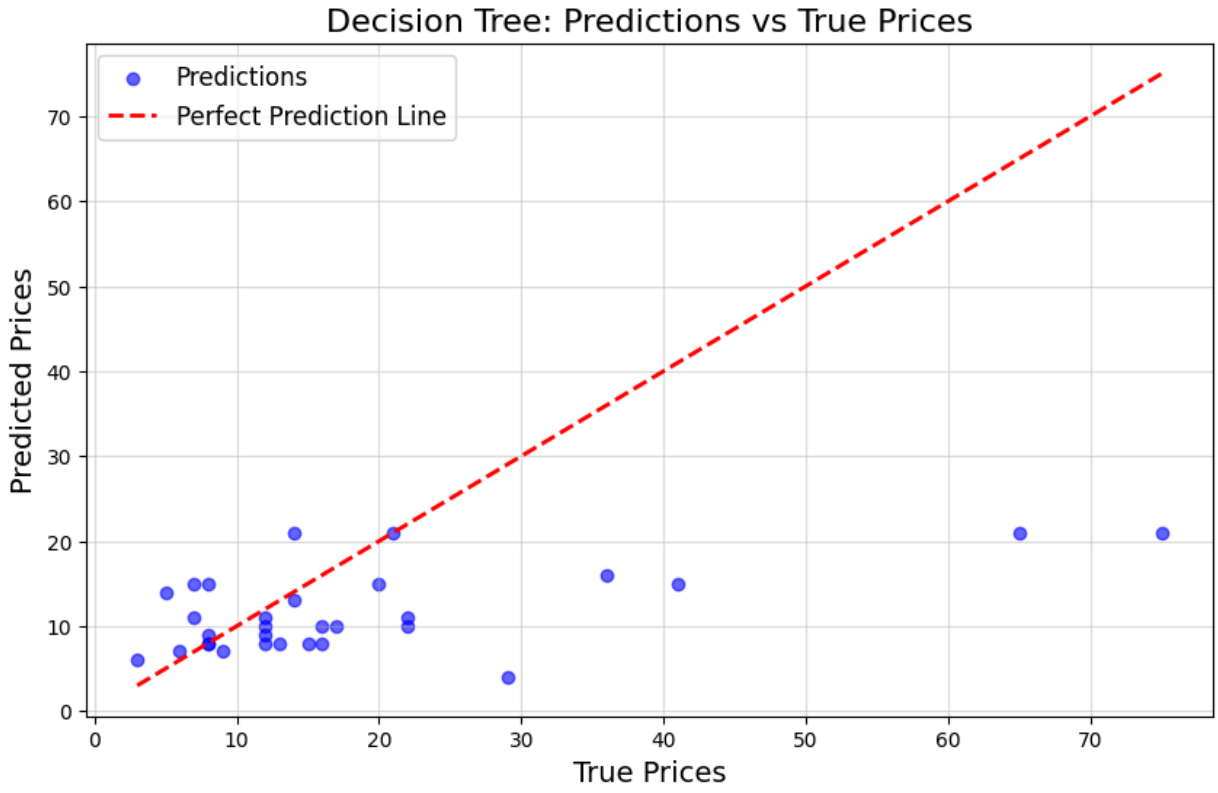
In order to optimize the Ridge Regression model, I modified the alpha parameter using RandomizedSearchCV. This resulted in an improved R^2 score from the Ridge Regression model of -0.1680. The results of the modified Ridge Regression model can be seen in Figure 4 below; this model shows improvement, but the R^2 value is still not ideal.

Figure 4:



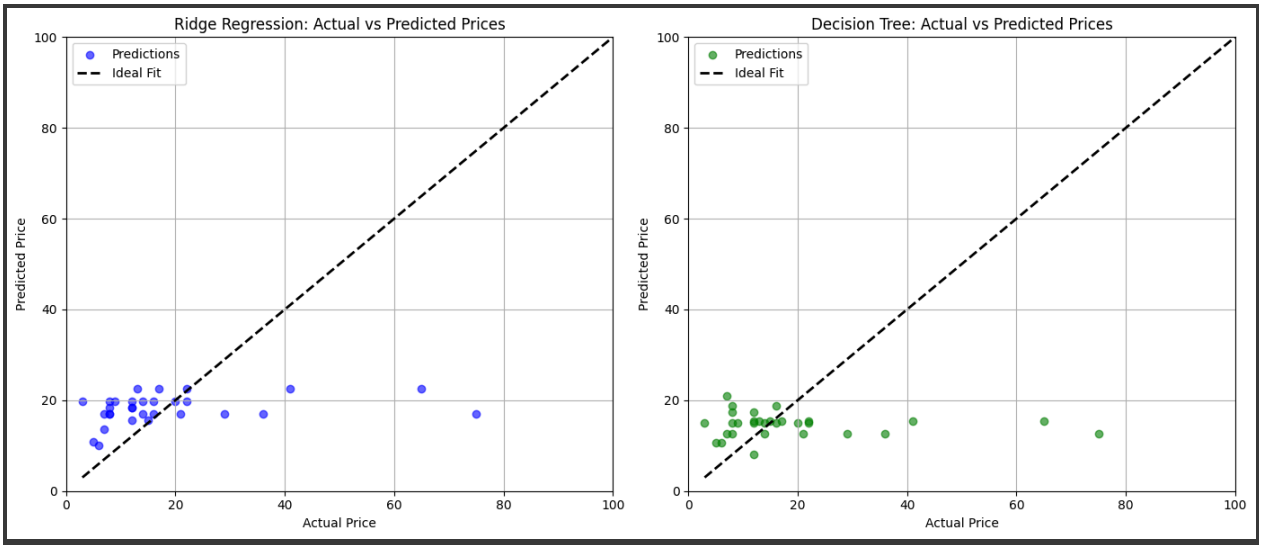
A Decision Tree model was employed next, without adjusting parameters. The R^2 score of this model was 0.072, and the model results can be seen in Figure 5 below. This is the best model so far, although at this point one can notice a trend of which entries the model is misclassifying. As stated previously, one common brand was an outlier in terms of price compared to the other most common brands in the set. Without analyzing by brand, the points that the models are misclassifying in terms of expected price, are those with high prices. This is likely due to the fact that many of the entries in the data set all had similar resale prices of around $17, making it harder for the model to identify features that cause the few high resale values in the dataset.

Figure 5:



In an attempt to assist the model with this challenge. I tested both the Ridge Regression and Decision Tree models on a subset of the data that contained only the top 10 most common brands. I then gave each a numerical ranking from 1-10 (with 1 being the most expensive) based on mean price. This replaced the one-hot encoding approach to brands above and neither model contained specific modifications. The R^2 scores of this new approach were 0.0738 and -0.1125 for the Ridge Regression and Decision Tree models respectively. The results can also be seen in Figure 6 below. This strategy led to a worse outcome for the Decision Tree Model, but improvement for the Ridge Regression Model. This could be because the Ridge Regression model benefits from the regularization applied to handle the sparse and high-dimensional nature of the data, which aligns well with the challenges posed by one-hot encoding. The Decision Tree model's decline in performance with the new encoding strategy could stem from its sensitivity to sparse data and its tendency to overfit smaller subsets of data when not adequately tuned.

Figure 6:



Discussion

The ability of the models to accurate predict the resale value of the samples in this dataset are be ranked from best to worst based on their R^2 scores:

- Ridge Regression (ranked top brands)
- Decision Tree
- Decision Tree (ranked top brands)
- Ridge Regression (optimized alpha)
- Linear Regression

Some models performed better than others because of their ability to handle specific data challenges. Ridge Regression outperformed Linear Regression due to its regularization capability, which helped manage the sparsity introduced by one-hot encoding and reduced overfitting. Decision Trees, while well-suited for capturing non-linear relationships, struggled with overfitting in the presence of sparse, categorical data. Additionally, their performance suffered when encoding strategies, like ranking top brands, altered the structure of the data, highlighting their sensitivity to input transformations.

The variability and complexity of the dataset also played a role. The similarity in resale prices for most items and the presence of outliers made it difficult for all models to accurately predict high-value items. Brand and description features were particularly difficult to handle because of their variability. Natural language processing (NLP) techniques would likely improve the handling of item descriptions, but these methods are computationally expensive and outside the scope of this project. Additionally, the dataset includes a wide variety of brands, which complicates attempts to rank or categorize them effectively.

Future work could explore more robust methods for handling brand and description features. For brands, clustering techniques could group similar brands based on pricing patterns, reducing dimensionality and improving the model's interpretability. For item descriptions, implementing NLP techniques could extract meaningful information to enhance prediction accuracy. Expanding the analysis to include other categories or increasing the dataset size could also improve model performance by providing more data to identify patterns and trends.

Conclusion

This project explored the use of machine learning models to predict the resale value of blouses in the "Women/Tops & Blouses/Blouse" category, offering insights into the challenges and opportunities of using such techniques for sustainable fashion decisions. Ridge Regression, especially with top brand ranking, performed the best due to its ability to handle high-dimensional data and mitigate overfitting. While Decision Trees captured non-linear relationships, they were sensitive to sparse and transformed data. Key findings highlight the difficulty of modeling resale values given the dataset's variability and the importance of preprocessing strategies. Future work could focus on integrating NLP for item descriptions, clustering techniques for brands, and expanding the dataset to improve prediction accuracy and support environmentally-conscious decision-making.