Machine Learning Component

Model Selection

We chose to use Linear Regression and KMeans Clustering for our ML component. We chose Linear Regression because we wanted to predict the price of items, which is a regression task. Additionally, we chose KMeans Clustering to see whether there are patterns in the data that would naturally reflect the distribution of prices. Using KMeans Clustering, we also use different subsets of attributes to qualitatively see whether price distribution between clusters would differ.

Metric Selection

For Linear Regression, we used Root Mean Squared Error and R-squared score to measure the success of our model. We chose Root Mean Squared Error because it is in the same unit as the target, which is in US dollars, giving us a direct comparison of the difference between our predictions and the ground truth. We also chose R-squared score because we also wanted to know if our model was able to predict results that correlated with the ground truth as mean squared error could be misleading by averaging varying values.

Challenges

Some challenges we faced was scoping the data such that it reflects realistic use cases of our model. We realized that our data contained outlier items that were either priced to be free or the maximum possible price value unseriously, both of which we reasoned were cases beyond the scope of our project because people who want to sell something for free would not need a model to predict the price and people pricing something as the maximum price would not have done it seriously. Therefore, we removed the items with prices below the 5th percentile or above the 95th percentile.

Data Restructuring

We preprocessed our item title text data using the small text embeddings from SpaCy. Additionally, we preprocessed our image data using the VGG16 Keras model trained on ImageNet, however since the result was 1000 dimensions, we used Principal Component Analysis to reduce the dimensions of the image features to 210 dimensions, which retained 90% of the explained variance.

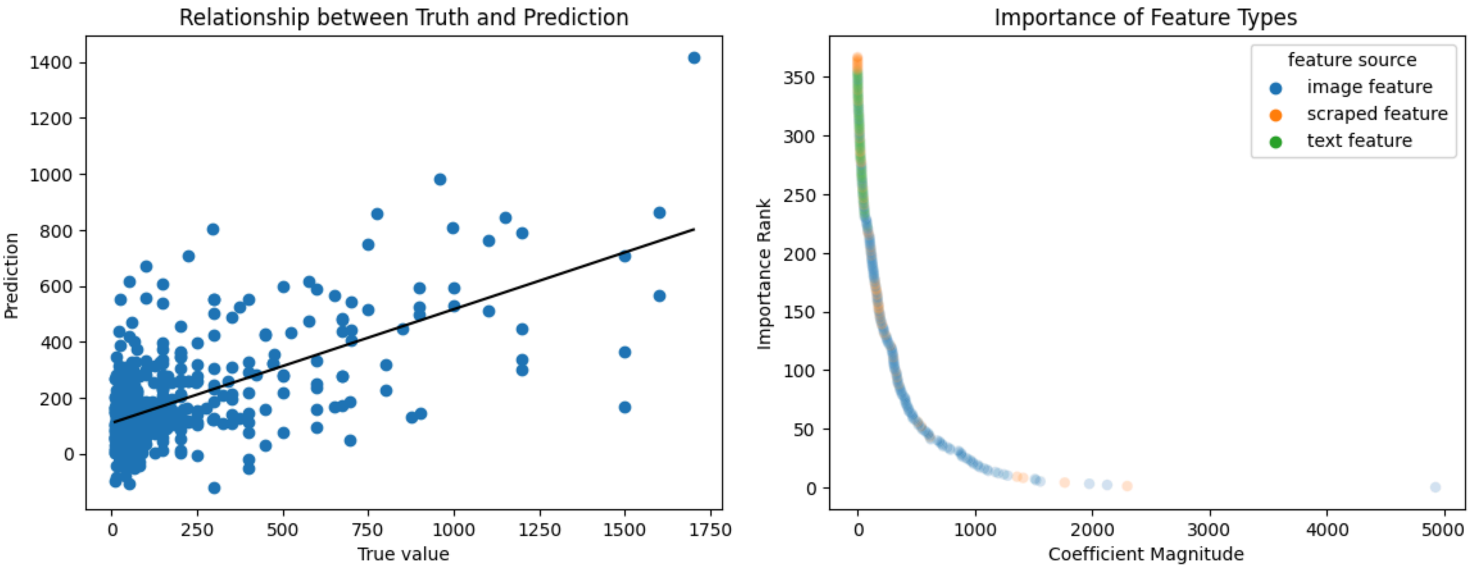
Results

For Linear Regression, we ran Cross Validation with 5 folds of training and testing data. In terms of metrics we use RMSE and R2 to measure prediction success.

For KMeans Clustering, we measure success qualitatively using histograms. We compare the price distribution based on clusters from different subsets of attributes.

Trial	**RMSE**	**R2**
1	256.99	0.1254
2	262.70	0.0383
3	271.00	0.0802
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.



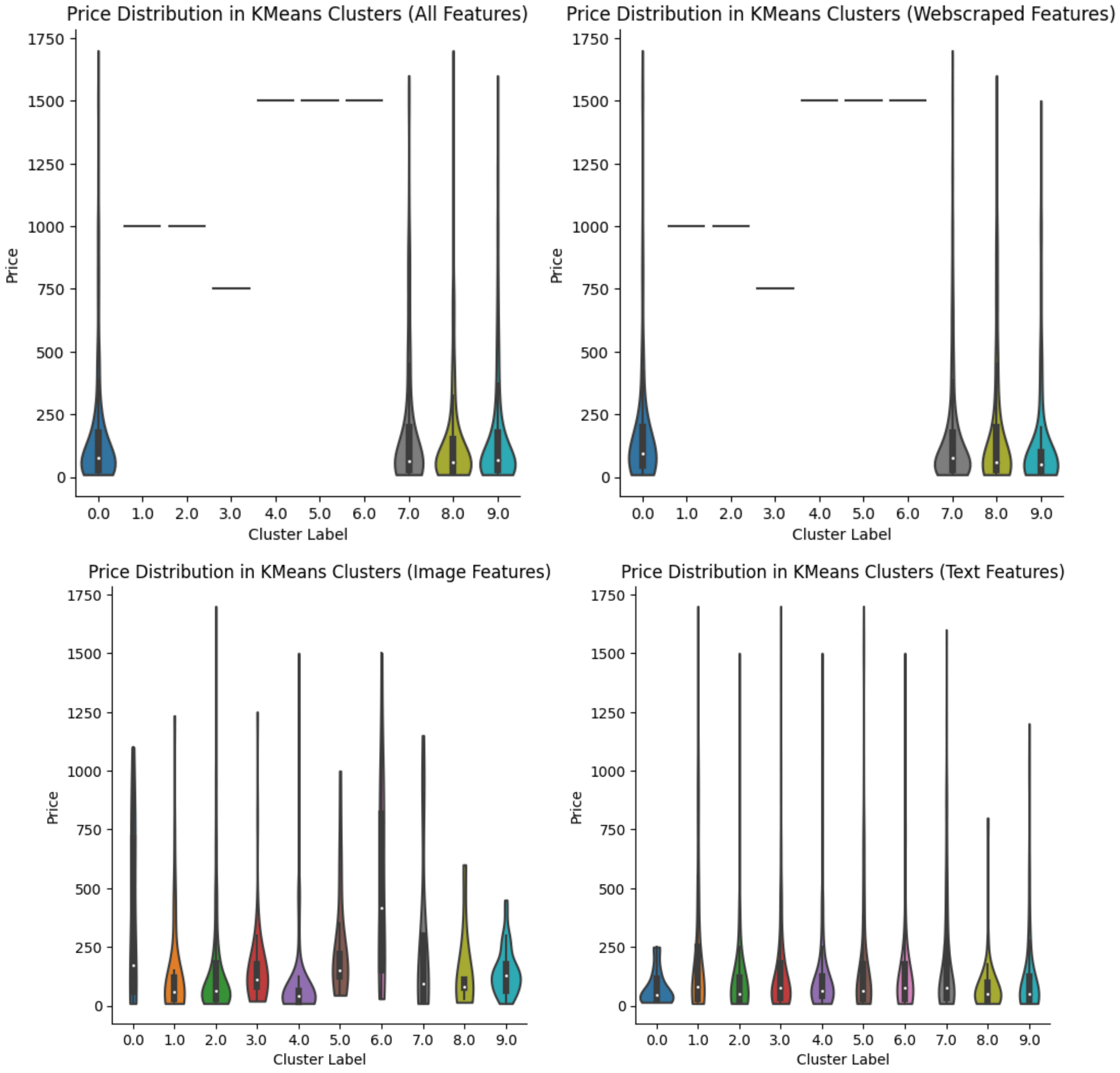
Graphs: (Left) Correlation between Ground Truth and Predictions. (Right) Importance of different feature types. The lower the point, the more important the feature.

Linear Regression

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.

In the qualitative results, we see from the left graph that the line of best fit between ground truth and prediction does trend upwards, which signals a positive correlation. This reinforces the quantitative results. In the right graph, we see that the features with the largest coefficients are either image or webscraped features. Among the largest webscraped features are dummy variables that represent whether the item is an automobile, which makes sense as automobile items are generally more expensive.

In both quantitative and qualitative results, it seems that the features have a slight positive correlation with the prediction, however due to the large root mean squared error for price, I would not trust this model to make a very accurate price prediction.


Price Distribution in KMeans Clusters (All Features)

Price Distribution in KMeans Clusters (Webscraped Features)

Price Distribution in KMeans Clusters (Image Features)

Price Distribution in KMeans Clusters (Text Features)

Graphs: Price Distributions in KMeans Clusters clustered from different features subsets. (Top Left) All features. (Top Right) Webscraped features. (Bottom Left) Image Featuers. (Bottom Right) Text Features.

KMeans Clustering

From the qualitative results, we see that there are certain webscraped features that influence the clustering algorithm to produce outlier one-sample clusters based on the top left and top right graphs. This is likely due to the presence of categorical or dummy variables in the subset of webscraped features such as “item category”. Disregarding the outliers, it seems that there is little difference in price distribution across clusters. The only clusters that do have changes in price distribution is the subset of image features, which is expected given the magnitudes of feature coefficients shown in the qualitative results for Linear Regression.

Discussion

Due to the small R2 score and high RMSE score in our LInear Regression, these predictions should not be used decisively. However, given our use case of providing a price suggestion for an item, we can afford a higher margin for error since these prices should still ultimately be decided by the seller. Therefore the Linear Regression model may still be a good model for providing price suggestions.

Additional Comments

These results do correspond with our initial belief in the data. Since the price can really depend on so many nuances such as item condition, type, category, location, etc, traditional Machine Learning models may not be best suited for processing such complex combinations of data types. Additionally, depending on the person and their emotional attachment to items, even the same item with similar conditions may still be valued differently. These are attributes that are especially hard to estimate or take into account.

I believe that the tools for analysis chosen are appropriate. For the hypothesis tests, we aimed to separate the overarching price prediction goal into smaller subsections. And for machine learning components, we aim to see whether this overarching task can even be achieved with traditional machine learning techniques that are more explainable. Therefore, in the context as a proof of concept, I believe that the modelling technique of Linear Regression and analytical technique of KMeans Clustering is sufficient to reveal the challenges and possibilities of this problem.

I believe that the data is adequate for our analysis and goal. Since our goal is to see whether information provided by a seller before inputting the price is sufficient to give a decent prediction of the price, we collected all the information that would be available when deciding on a price. Therefore, the data collected was definitely an adequate representation of the problem scope that we aim to tackle.