ORIE 5741 Final Report

Catherine Appleby, Younes Bensouda, Louis St-Pierre

Table of Contents

Introduction

The goal of this project was to produce a more reliable system to price art than what is currently available. Often in auctions, art pieces sell for well outside of the bounds experts give on the piece's value. Knowing why or being able to predict when this will happen could provide more stability for art as an investment, and could help to protect buyers, sellers, and auction houses when they put a piece up for auction. Thus, we aim to build a model that can reliably predict the price of a piece of art based on available information before bidding in an auction starts. We used data from the study "Death, Bereavement, and Creativity," which includes art auction records for contemporary and impressionist paintings ranging from 1972 to 2014 (Graddy, 2017). The dataset has around 14,000 records, with 34 features including data on the artists, qualitative information about the paintings, and the auction houses of each auction.

Preprocessing

The first step of data preprocessing was to look at all data formatted as text and not as real numbers. This included artist and painting names, auction house names, and important dates. Dates were transformed into real number values or removed if they were redundant. The auction house names, and painter names were formatted as one-hot vectors, since they are not ordinal, and they are mutually exclusive and independent. However, with 294 auction houses in the original dataset, this would significantly increase the number of features and we were concerned of the potential of overfitting. However, as seen in Figure 1, the majority (78%) of painting sales occurred in just 3 auction houses. Thus, we dropped all sales not in the top four auction houses.

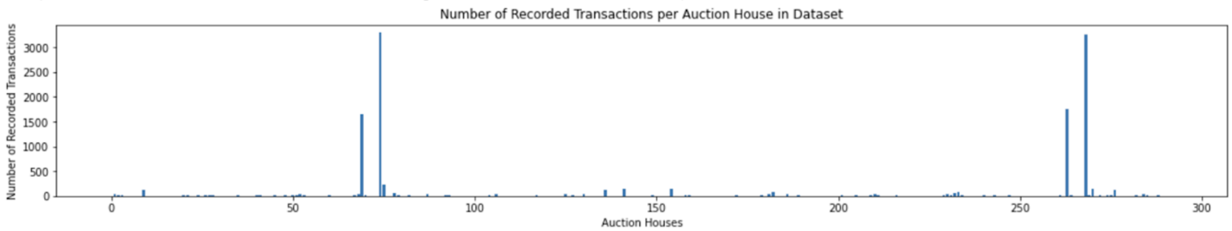


Figure 1: Sales per auction house

We conducted a similar analysis on the artists in the dataset, but the sales per artist were much more evenly distributed and did not justify such filtering (Figure 2). Furthermore, tests were done in the ridge regression stage to see if filtering out artists with fewer sales would increase model performance, and it had no effect.

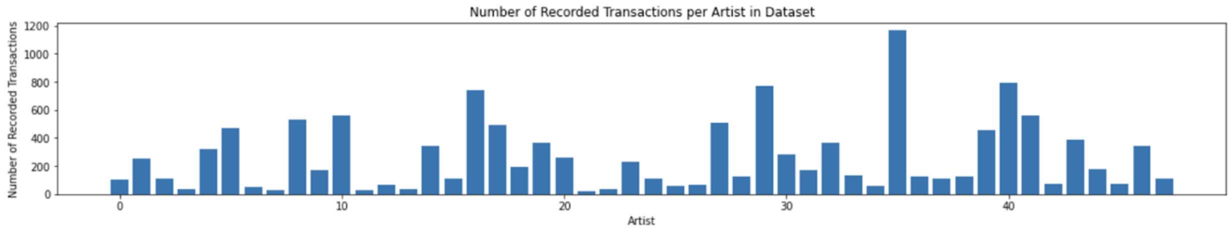


Figure 2: Sales per Artist

The next step was to drop columns that were either redundant or held information that we do not want our model to have access to. Redundant columns include columns where dates appear several times, and when measures appear in different units. The information which a seller does not have access to before the auction process is the auction lot number, whether the sale was at auction hammer or not, the consumer price index for the given year relative to a future date, and most importantly, the appraised value of the painting. We want our model to not depend on price estimates since we want it to be scalable and easily applicable, and to not depend on a costly, lengthy, and sometimes inaccurate appraisal process.

Regarding missing values, Figure 3 displays the number of missing values in the original dataset. The columns holding the year of death for relatives show many missing values. From our judgment, lacking information on the death dates of relatives for certain artists was not judged to be of significant relevance to our estimation objective of sale price, and so we filled the missing values with 0. There were many missing values in the expert price estimates columns, but these columns will be dropped from the model as mentioned previously. The rows missing sale price values were dropped, since these would be required for training and validating our model using supervised learning methods. After these steps, no missing values remained.

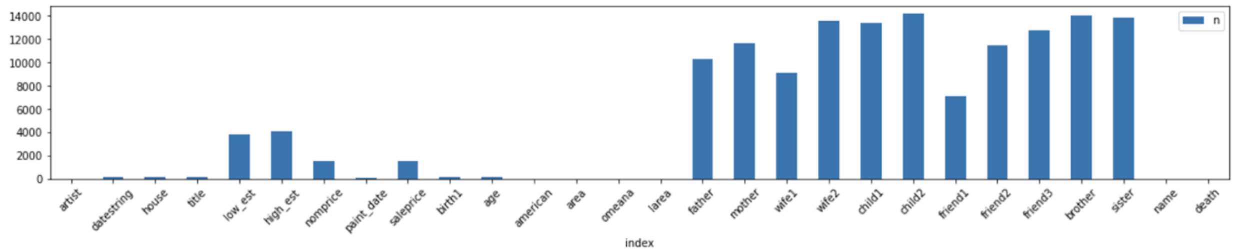


Figure 3: Prevalence of missing values in original dataset

The last preprocessing step involved transforming the sale price of paintings to the log price of sale. This is because the sale prices were heavily skewed to the right. With such a heavily skewed distribution, there was a concern that models would focus exclusively on the high-priced paintings, disregarding the bulk of the dataset. The log prices are much more symmetrically and normally distributed, which should facilitate interpretation and modeling.

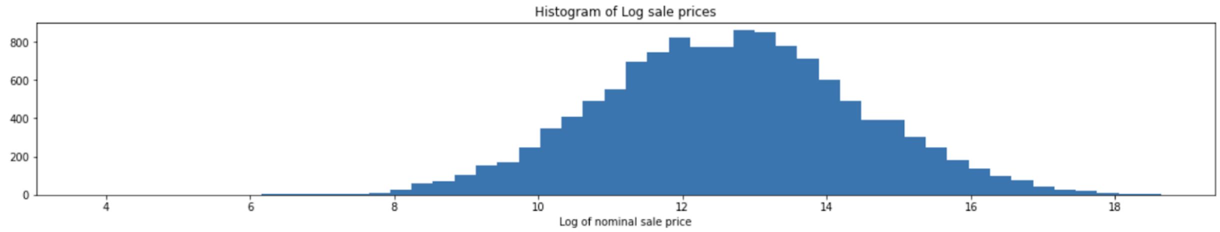
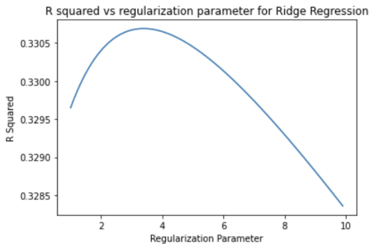
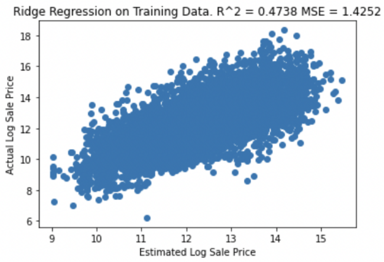
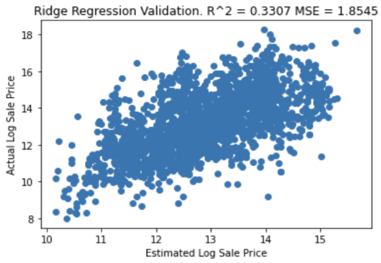


Figure 4: Distribution of log prices

Regression and Data Addition

After all the preprocessing was completed, we have a dataset that is completely real-valued, with anomalies and missing values removed. This dataset is now ready to be used for regression modeling. First, we split the data into train, validation, and test sets, with sizes of 60%, 20%, and 20% of the original dataset, respectively. We sorted the sales by date of sale, to mimic a time-series dataset, otherwise information of previous sales could be inferred from the training dataset based on similar dates, artificially improving test performance. We applied linear, lasso, and ridge regression on the preprocessed dataset and found ridge regression to be the most robust and which produced the best results. The regression process involved running ridge regression on the training set over a range of regularization parameters, selecting the best parameter, and then using that parameter on the validation set.

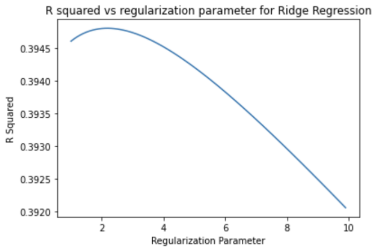
Table 1: Ridge regression on original data

Grid search on Training Set	Training Results	Validation Results
		
Parameter = 3.75	MSE = 1.43	MSE = 1.85

The analysis did not stop there, however. We determined that when evaluating price, there were two important features which we could add which would improve model performance and could be added without cheating. Firstly, for each row, we can add a feature that includes the price of the last painting sold by the same artist. This is information available to people at the time of desired sale and respects the rules of temporal information. Secondly, in the same vein, for each row of the dataset, we can add another feature that includes the last sale price of that specific painting. So, for a potential sale, we have added information on the last painting sold by the same painter, and the previous sale price of the specific painting up for sale. There were also "no-data" columns for each of these features, to capture the information implied by being the first time on record that the painting or the artist was making a sale. Running the ridge regression on this enhanced dataset improved the results, as seen in the table below.

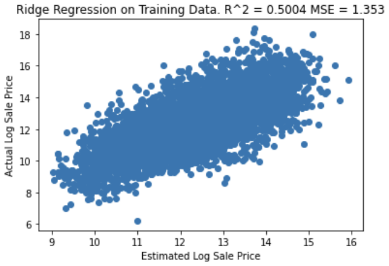
Table 2: Ridge regression on enhanced data

Grid search on Training Set



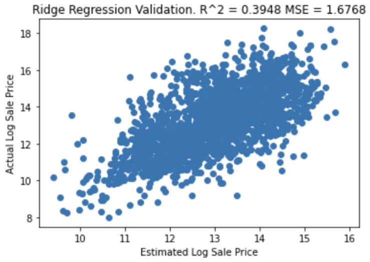
Parameter = 2.25

Training Results



MSE = 1.35

Validation Results



MSE = 1.67

Lastly, in an additional attempt to improve results, we vectorized the titles of the artworks and included the vector as a feature. The resulting model would simply overfit on the additional features and did not perform better in the validation stage. It was concluded that information captured by the painting title, if at all present, could not be adequately exploited.

In conclusion, the ridge regression model was tested on the enhanced dataset without title vectorizations, and the test MSE was found to be 1.8 on the log prices. The mean absolute error for prices was 2.33 million. Due to the benefits from using the enhanced dataset, it will be used for all subsequent models.

Support Vector Regression

SVR is a variant of Support Vector Machines (SVM) for predicting discrete values. We wanted to use it because SVM is usually effective in high dimensional spaces. The SVR here tries to fit the best line within a threshold value, and it depends only on a subset of the training data, because the cost function ignores samples whose prediction is close to target. However, the performance of this model in our study is not satisfactory. We get a MSE on validation data of 3.36, and 3.98 on test MSE, and this is not interesting compared to the other methods considered.

AutoML

AutoML is an automated toolkit to provide guidance on which model could perform well on a dataset. After applying it to our dataset, it recommended the investigation of tree-based methods. Specifically, Random Forest was the top model it recommended. We investigated other methods based on intuition, but it is of interest to note that this was the recommended method based on AutoML, and that the Random Forest method was proven to have the lowest Mean Squared Error on the validation set of all the models we fit.

Clustering

To know more about our data, we also decided to make a clustering analysis using K-Means Clustering. It was necessary to reduce the dataset to two dimensions using PCA to be able to visualize the clusters of observations. With this method, we could retain 77% of the variance with two components.

The optimal value of k can be found with the elbow method. For that we run k-means with each number of clusters, and we compute the sum of squared distances to the centre of each cluster divided by the inertia. This leads to Figure 5, which seems not to be perfectly elbow-shaped here, but we decided that the best value of k was 8.

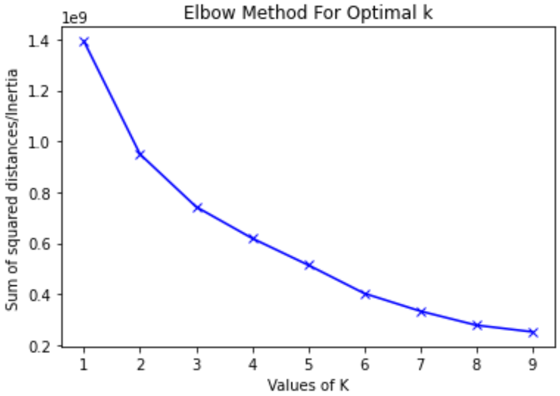


Figure 5: Elbow method for optimal k

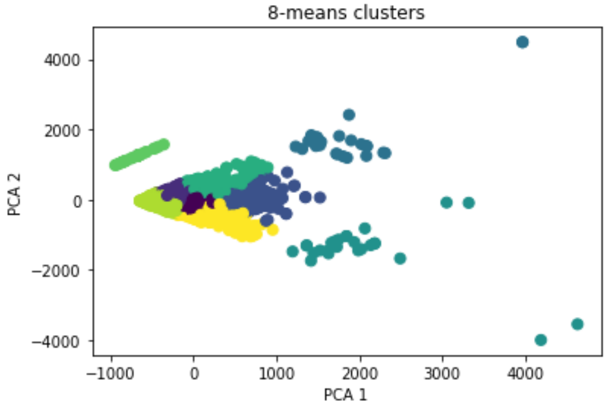


Figure 6: K-Means clustering visualization

The K-Means algorithm tries to cluster data by separating groups of data on k groups of equal variance, minimizing the inertia, that is the sum of the distance between the center of a cluster and each point belonging to this cluster.

Because K-Means works well with convex and isotropic clusters (typically spheric), K-Means may not work well with elongated clusters or irregular shapes. Here, the shapes of our clusters are not exactly spheric, and sometimes elongated, but we believe this is an interesting approach for clustering here.

When analyzing each cluster, some groups of artists are appearing, and are somehow referring to actual artistic movements, confirmed with the other features.

Clusters 0 and 7 show a large number of art pieces from Picasso and Chagall, who both well represent Cubism and Surrealism. In cluster 1, Léger and Motherwell indicate a group of expressionists and modern artists. Monet and Picasso in cluster 2 show both impressionism and post-impressionism trends, as well as in cluster 3 with Renoir. Cluster 4 is mostly representative of surrealism and abstract expressionism with Tanguy and de Kooning. Finally, in cluster 5 and 6, we can find Monet and Vuillard, both impressionists.

Regression Tree

The data lends itself well to a regression tree, as we are aiming to identify similar paintings and sales to inform on the price of subsequent auctions, thus we want 'bins' of similar sales prices for similar paintings. Each node in this tree splits based on whether the value of a feature is above or below a split that reduces mean squared error. Since this is a regression tree and there are a large number of features, the depth of the tree needs to be set so that it does not overfit on the training data. Each leaf, or terminal node, on the tree represents a group of paintings that are priced similarly, and provides a mean price for every painting in that group. To prevent overfitting, we tested multiple different depths, using the DecisionTreeRegressor from sklearn, as seen in Figure 7. We picked the depth based on the minimum mean squared error produced by the validation dataset, which was 2.19 at depth 5. Though not used as a metric for choosing the model, we see this is also close to a minimum of the test mean squared error. Figure 8 shows a regression tree fit to a depth of 2, resulting in low mean squared errors on the training dataset.

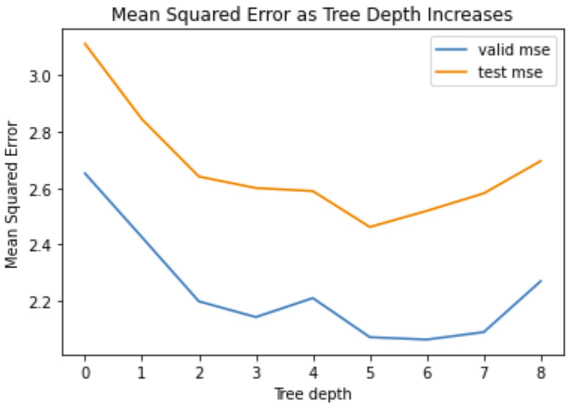


Figure 7: MSE as tree depth increases

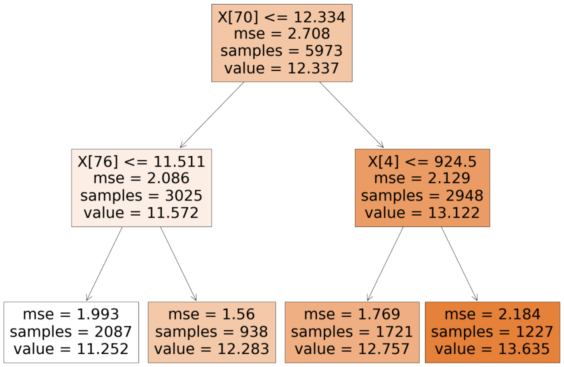


Figure 8: Tree visualization

To improve on the regression tree, we first applied AdaBoost, to attempt to reduce errors. Using a maximum depth of 5, the model did improve performance on the validation dataset compared to the single decision tree, with a mean squared error of 1.76.

Random Forest

Ensemble methods are an intuitive improvement on a single regression tree, especially since we have such a large number of features. However, AdaBoost didn't perform as well as Ridge Regression, thus we tried Random Forest, another ensemble method. Note that AutoML also suggested the use of Random Forest. We used a default number of 100 trees when running the RandomForestRegressor from sklearn, though we tested different maximum depths to find a model that performed well, but didn't take too much computation time, resulting in a chosen maximum depth of 25. This resulted in a validation mean squared error of 1.63, the best validation error of all the models.

Results

Table 3: MSE by method

Model	**Best validation MSE**	**Test MSE**
Ridge Regression	1.7	1.8
SVR	3.36	3.98
Regression Tree	2.19	2.59
AdaBoost	1.76	2.14
Random Forest	1.63	2.02

The metric used to evaluate our models was mean squared error (MSE) on the log-prices. The model with the best MSE on validation data was the Random Forest model. Thus, we consider Random Forest as the best performing model in this study, a conclusion which is supported by the recommendations of AutoML. However, the lowest MSE on the test dataset is from Ridge Regression. We include this as a point of interest, but we still consider Random Forest as the best model since the test dataset cannot be used to make decisions on model selection or parameters.

Discussion and Conclusion

In conclusion, we do not recommend using our model for real-world applications. The absolute error for our best model - the random-forest model - is far too large and there is no evidence that it effectively estimates sale prices, despite numerous efforts to enhance and capture information within the dataset. Supporting this conclusion is the diversity of models with which we applied to the data, all yielding similar results. It is particularly interesting how random forest and ridge regression both produced mean absolute errors of around 2.3 million in the real price space (not log-price).

Regarding the fairness of the models we constructed, we believe the principal concern would be that models would be highly tuned to artists fitting a certain background. In our dataset, the artists were American or European. The concern would be using this model to evaluate artists from different backgrounds, or who specialize in different media, and perhaps undervaluing them. This undervaluation of their art could have real impacts on their careers. The risk of creating a “Weapon of Math Destruction” is moderate. If this model was implemented for price appraisals, it is possible the price appraisals are used as a grounding value for bids and could influence auction outcomes. However, since the participants in the art market are diverse and heavily incentivized to exploit vulnerabilities in pricing models, weakness in pricing fairness could be arbitraged out. Further analysis would need to be done to see how appraisals affect the outcome of auctions and if there is a significant feedback effect.

Suggestions for improving this project would be to collect more data from a more diverse group of artists. Looking only at 14,000 sales of paintings from 46 artists does not come close to characterizing the world market for paintings, much less art in general. Also, broadening the scope of analysis to include more markets than simply 4 auction houses could improve performance, as a greater share of the art market could be captured and analyzed. Lastly, adding data about contemporary financial and economic conditions could potentially improve estimates, as these can incentivize investment towards or away from the art market.

References

Kathryn Graddy, Carl Lieberman (2017) *Death, Bereavement, and Creativity*. Management Science. Published online in Articles in Advance 16 Oct 2017.
https://doi.org/10.1287/mnsc.2017.2850