Final Data Analysis I

Yaoyao Fan, Lucie Jacobson, Zining Ma, Jiajun Song

2019-12-07

Summary.

We are four art consultants analyzing the prices of auctioned paintings in Paris from the years 1764 to 1780. The principal objective of our analysis is to predict the final sale price of auctioned paintings in 18th century Paris, identifying the driving factors of painting prices and thereby determining instances of under- and over-valuation.

Data.

The data utilized in the analysis is provided by Hilary Coe Cronheim and Sandra van Ginhoven, Duke University Art, Art History & Visual Studies PhD students, as part of the Data Expeditions project sponsored by the Rhodes Information Initiative at Duke. To begin, there are three subsets of the complete data set - one subset for training, one subset for testing, and one subset for validation. The training subset, which is utilized during exploratory data analysis and initial modelling, is comprised of 1,500 observations (paintings) of 59 variables that provide information pertaining to the origin and characteristics of the artworks.[1]

[1] Detailed descriptions of all variables are available in the attached MD file, `paris_painting_codebook.md`.

Research Question.

What are significant predictors for the final auction sale of a given painting in Paris from the years 1764 to 1780? Is the resulting statistical model diagnostically adequate for the prediction of the sale price for a given painting?

Why Our Work is Important.

"Speaking in the most basic economic terms, high demand and a shortage of supply creates high prices for artworks. Art is inherently unique because there is a limited supply on the market at any given time" [2]. Indisputably, art is extremely important across cultural and economic spheres. Art history provides exposure to and generates appreciation for historical eras and global culture, and thus correct art valuation provides a standard metric for both the trained and the untrained eye to distinguish amongst historical artworks, consequently influencing the framework of modern art as well.

[2] referenced from "Art Demystified: What Determines an Artwork's Value?", available at `https://news.artnet.com/market/art-demystified-artworks-value-533990`

Exploratory Data Analysis.

Using EDA and any numerical summaries, get to know the data - identify what you might consider the 10 best variables for predicting `logprice` using scatterplots with other variables represented using colors or symbols, scatterplot matrices or conditioning plots.

Response Variable.

To begin, we analyze the selected response variable, `price`, and the log-transformation of `price`, to ensure that the response variable is approximately normally distributed.

From *Figure 1*, we observe that the distribution of the variable `price`, with range from 1 to 29000 (note: 1 livre sterling is approximately equal to $1.30 U.S. dollars), is strongly skewed to the right. This is corroborated by the normal probability plot for the data, which fails to conform to a linear trend. This is expected, as it is reasonable to assume that on average, prices of paintings at auction will fall within a reasonable budget range: the entire range, however, has a lower bound greater than 0 and potentially no upper bound - the price can be whatever an individual is willing and able to pay for a particular painting.

Given the histogram for `price` is strongly skewed, we now consider the log-transformation of the variable. Logarithmic transformation is a convenient means of transforming a highly skewed variable into a more closely normally-distributed variable, and this transformation is commonly used in economics and business for price data.

The histogram of the variable `logprice` now exhibits significantly less skew, and much more closely approximates the normal distribution. We also observe that the normal probability plot for the data follows a general linear trend, except in the tail areas of the distribution. We conclude that the conditions for inference regarding the distribution of the variable of interest are sufficiently met, and we continue with the exploratory data analysis.

Data Manipulation.

To begin data manipulation, we divide variables based on their data type and analyze them.

We first deal with all character variables. We observe that the variable `lot` should be numeric. We then determine which character variables should be categorical factor variables, where the number of unique levels is restricted to less than 15[3](this is an arbitrary cut-off point, but is necessary - variables with too many levels will not have enough observations in every level to generate robust estimates). To

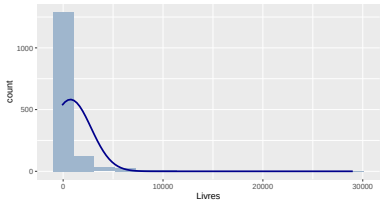


Figure 1: Histogram of Painting Price Fetched at Auction (Sales Price in Livres)

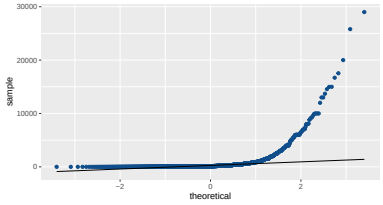


Figure 2: Normal probability plot of Painting Price Fetched at Auction (Sales Price in Livres)

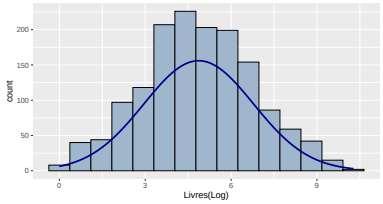


Figure 3: Histogram of Log Painting Price Fetched at Auction (Sales Price in Livres)

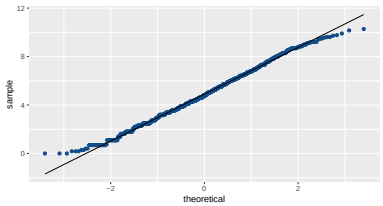


Figure 4: Normal probability plot of Log Painting Price Fetched at Auction (Sales Price in Livres)

[3] We omit variables `sale`, `subject`, `authorstandard`, `material`, `mat` at this step. Note that there might be useful information in these variables therefore further attempt on improving the model may require take a look at them.

initially handle "NA" and blank observations, we:

- impute a value of "Unknown" to all "n/a" variables for `authorstyle`,
- a value of unknown ("X") to all blank observations for `winningbiddertype`,
- a value of unknown ("X") to all blank observations for `endbuyer`,
- a value of "Unknown" to all blank observations for `type_intermed`,
- a value of "Other" to all blank observations for `Shape`, and
- a value of "other" to all blank observations for `materialCat`.

DataType	*Count*
character	17
categorical	10
continuous	32

Our initial data analysis reveals that there are 7 unique levels for the variable `Shape`. We observe that two levels are "round" and "ronde", and two levels are "oval" and "ovale". We learn that "ronde" is the French word for "round" and "ovale" is the French word for "oval", and thus we combine observations in the respective levels. The resulting levels are: "squ_rect", "round", "oval", "octagon", "miniature", and "Other".

Similarly, multiple levels of the variable `authorstyle` are quite similar: "in the taste of", "in the taste", and "taste of": thus, we group all of these unique levels into one level, "in the taste of". A summary table of the character variables is presented below.

We then coerce all variables in the character type data frame to be of type factor.

Summary of All Initial Character Variables. Note that here X and Unknown both stand for missingness or data not available. Such imputation may lead to bias in prediction. We should be careful with these variables.

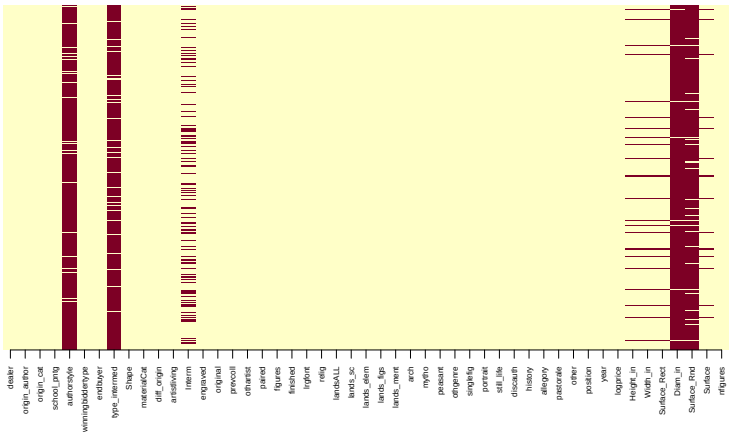
dealer	origin_author	origin_cat	school_pntg
J:201	A : 7	D/FL:594	A : 1
L:263	D/FL:590	F :483	D/FL:658
P: 93	F :578	I :170	F :608
R:943	G : 26	O :251	G : 1
	I :159	S : 2	I :193
	S : 11		S : 2
	X :129		X : 37

authorstyle	winningbiddertype	endbuyer	type_intermed	Shape	materialCat
Unknown :1417	D :464	B: 14	B : 11	miniature: 2	canvas:731
after : 26	X :395	C:326	D : 94	octagon : 1	copper:131
in the taste of : 19	C :189	D:470	E : 39	Other : 20	other :229
copy after : 10	U :168	E:127	EB : 1	oval : 19	wood :409
attributed to : 7	E :127	U:168	Unknown:1355	round : 30	
in the manner of: 7	DC : 89	X:395		squ_rect :1428	
(Other) : 14	(Other): 68				

Missing Data.

We now identify factor, continuous, and discrete numeric variables, and generate a large data frame with all variables coerced to appropriate type. Let us determine which variables have unknown and/or missing data:

Figure 5: Determining NA Observations in the Data



From *Figure 5*, we observe that the variables `authorstyle`, `type_intermed`, `Interm`, `Height_in`, `Width_in`, `Surface_Rect`, `Diam_in`, `Surface_Rnd` and `Surface` all have unknown and/or missing data. We will analyze these variables further, beginning with `authorstyle`.

From *Figure 6* we can see that data is not missing at random, the missingness has something to do with our response. Thus we cannot just throw away observations and need to look deeper into these predictors.

From *Figure 7*, we observe that the majority of the observations for the variable `authorstyle` are "Unknown", with very few (or no) observations in the remaining levels. Consequently, this variable will likely not contribute much information for the prediction of `logprice` in any specified model, and the minimal number of observations included in the levels may generate extreme standard errors. Given this, we select not to include this term in model specification.

We will continue to analyze variables in the data set with significant numbers of `NA` observations.

Here, we observe that the majority of observations for `Diam_in`, the diameter of a painting in inches, and `Surface_Rnd`, the surface of a round painting, are `NA`. We note that the variable `Surface`, the surface of a painting in squared inches, effectively captures information for the size of a given painting. Including this variable in subsequent model

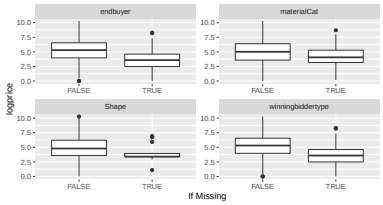


Figure 6: Missingness Effect on Response

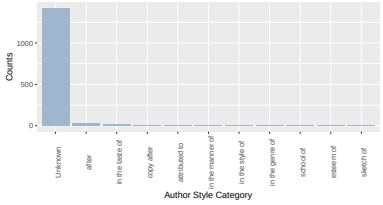


Figure 7: Counts of Author Style for Auctioned Paintings

specification captures information provided by the following variables: `Height_in`, `Width_in`, `Surface_Rect` and `Surface_Rnd`. Thus, we will include `Surface` in subsequent model specification and omit variables that are directly related to `Surface` to avoid issues of multicollinearity.

For "NA" values in `Surface`, we use the package "mice"[4] in R. MICE, Multivariate Imputation via Chained Equations, is considered more robust than imputing a single value (in practice, the mean of the data) for every missing value.

We now consider `Interm`, a binary variable that indicates whether an intermediary is involved in the transaction of a painting. This variable consists of 395 `NA` observations, 960 `0` (no) observations, and 145 `1` (yes) observations. Given this, we observe that many auctioned painting sales appear to occur without the involvement of an intermediary. This information is directly related to `type_intermed`, the type of intermediary (B = buyer, D = dealer, E = expert), and is only valid for the observations where an intermediary is involved in the transaction of a painting. Consequently, we select to omit `type_intermed` from the data set. However, we do note that the variable `intermediary` may provide information for the prediction of `logprice`, as *Figure 8* indicates that the median sale price for paintings where an intermediary is involved is noticeably higher than the median sale price for paintings where an intermediary is not involved. While the variability is quite high for both the "No" and "Yes" levels, the boxplot where an intermediary is not involved does not exhibit significant skew, while the boxplot where an intermediary is involved exhibits left skew.

We now look at information pertaining to painting material. We observe that there are initially 3 variables in the data set that pertain to painting material: `material`, `materialCat`, and `mat`. The levels of `material` are in French, and the English translations are precisely the levels of the variable `materialCat`. Additionally, we see that the variable `mat` is comprised of more levels (17, excluding "blank" and "n/a") than the variable `materialCat`, and thus is not included in our data frame (restriction of levels < 15). Let us determine if the variable `materialCat` lends information for painting price.

From *Figure 9*, we observe that the material category with the greatest number of observations is canvas, and the material category with the least number of observations is copper. However, the boxplot indicates that paintings with copper material maintain higher mean sale prices than paintings with canvas material; this may give evidence to the statement that "shortage of supply creates high prices for artworks".

Variable	*Number of Missing*
`Diam_in`	1469
`Surface_Rnd`	1374

[4] MICE is utilized under the assumption that the missing data are Missing at Random, MAR, and integrates the uncertainty within this assumption into its multiple imputation algorithm (referenced at `https://stats.idre.ucla.edu/wp-content/uploads/2016/02/multipleimputation.pdf`).

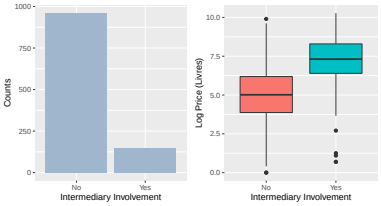


Figure 8: Painting Price and Intermediary Involvement

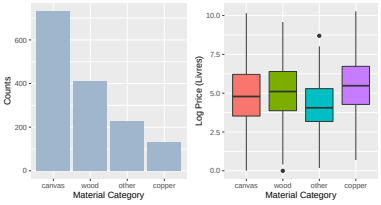


Figure 9: Painting Price and Material Category

Identification of Important Variables for the Prediction of Painting Price.

A boxplot matrix of selected variables of character type for subsequent model specification:

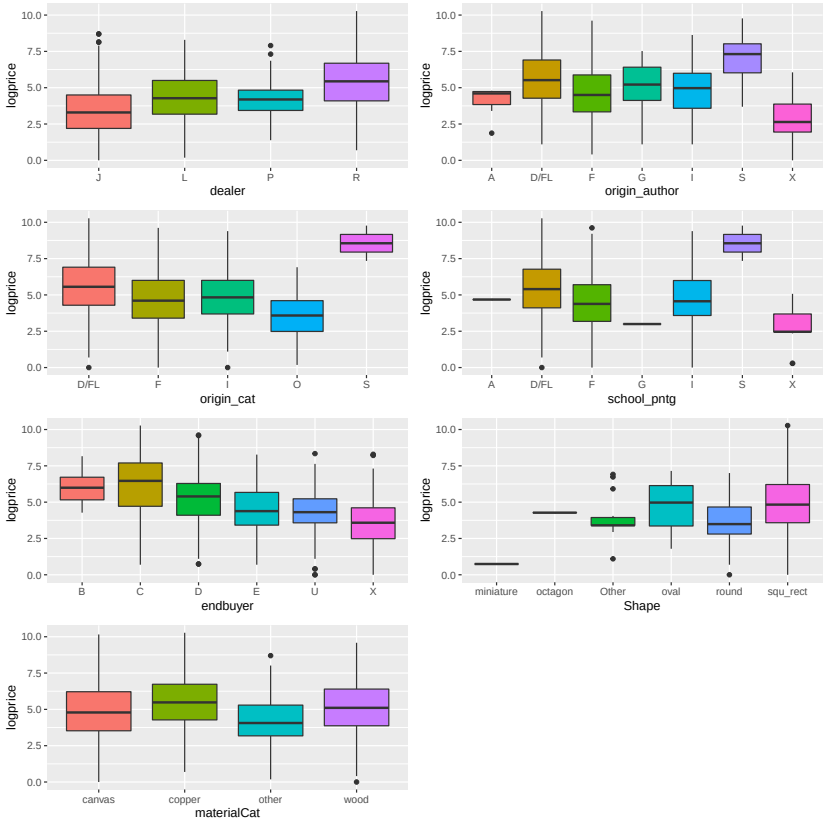


Figure 10: Boxplot of Character type predictors

From *Figure 10*, we note that different levels of `dealer` appear to have different medians of sale prices, with dealer "R" maintaining a higher median sale price than other dealers. We also note that paintings with Spanish author, origin classification, and school of painting appear to have noticeably higher median sale prices than other authors, origin classifications, and schools of painting (however, we know that there are limited observations pertaining to Spanish author and origin classifications in the data set, so this may not be a robust indication). Overall, all plots indicate trends within the variables that may be important for prediction of the auction price of paintings.

A boxplot matrix of selected variables of binary factor type for subsequent model specification:

As expected, observations that equal 0 for all binary variables do not contribute information for the auction price of paintings. We note

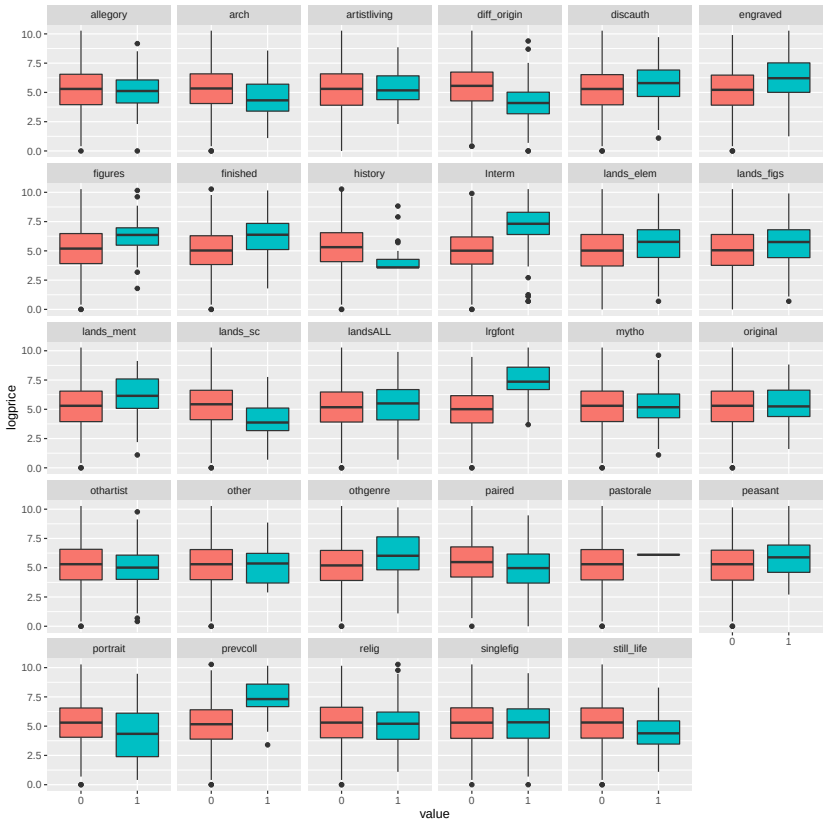


Figure 11: Summary Matrix for Binary Factor Variables

that the variables `lrgfont`, if a dealer devotes an additional paragraph (always written in a larger font size) about a given painting in a catalogue, `Interm`, if an intermediary is involved in the transaction of a painting, and `prevcoll`, if the previous owner of a given painting is mentioned, all have higher medians and higher price ranges with less variability than the other included variables. We also note that the variable `history`, if a description includes elements of history painting, appears to be associated with a lower median price on average.

A scatterplot matrix of the selected variables of continuous numeric type for subsequent model specification:

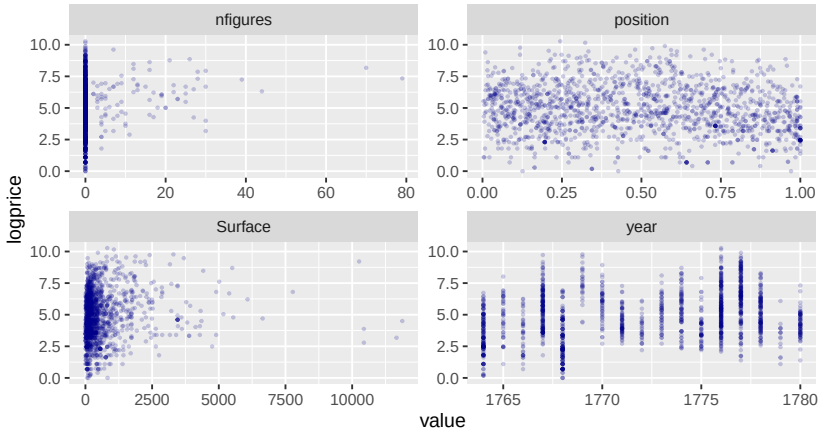


Figure 12: Scatter Plot Matrix for Continuous Numerical Variables

The variable `nfigures` refers to the number of figures portrayed in a given painting, if specified. Here, we observe that many paintings do not include any specified figures, and the prices for these paintings fall along the entire range of `logprice`. There is no clear trend for paintings that do include figures, and so we conclude that this variable is not likely to be very important for the prediction of the auction price of paintings. Similarly, the plot for `position` is a null plot with no trend. The plot for `Surface` indicates that there may be an association between the surface of a painting in squared inches and the price. Given the large range of the variable with several orders of magnitude, `Surface` should likely be log-transformed.

To further analyze potentially important predictor variables for `logprice`, we generate a random forest model. From the associated variable importance plot, we observe that the 10 variables resulting in the greatest increase in MSE are `year`, `Surface`, `endbuyer`, `origin_author`, `lrgfont`, `dealer`, `materialCat`, `paired`, `origin_cat`, and `Interm`.

Development and assessment of an initial model

Our final initial model is a linear model with these predictors:

`year`, `Surface`, `nfigures`, `engraved`, `prevcoll`, `paired`, `finished`, `relig`, `lands_sc`, `portrait`, `materialCat`, `year:finished`, `year:lrgfont` and `Surface:artistliving`

Development of model:

In EDA, we identified these variables to be important: `year`, `Surface`, `endbuyer`, `origin_author`, `lrgfont`, `dealer`, `materialCat`, `paired`, `origin_cat`, and `Interm`.

First we specified a large model, considering all 0/1 binary and factor predictors, as well as an interaction with `Surface` and `year`. An an interaction term implies that the effect of one of the given explanatory variables on the response variable varies for different levels of the second explanatory variable[5]. We developed our model this way because we inferred that there may be an association between `year`, `Surface`, and `logprice`, and under different conditions the linear relationship will be different.

[5] information referenced from "Interpreting Interactions in Regression", available at `https://www.theanalysisfactor.com/interpreting-interactions-in-regression/`

We then used the Akaike information criterion (AIC) for variable selection. The AIC is designed to select the model that produces a probability distribution with the least variability from the true population distribution[6]. While the AIC may result in a fuller model than the Bayesian information criterion (BIC) - which penalizes model complexity more heavily - the AIC criterion may lead to higher predictive power.

[6] referenced from "Akaike Information Criterion", available at `https://www.sciencedirect.com/topics/medicine-and-dentistry/akaike-information-criterion`

In our AIC model, there appeared to be several insigificant predictors. Thus, we decided to use Bayesian Model Averaging to extract the variables of the Highest Probability Model (HPM), and these variables were selected to be the variables in our final specified model.

Finally, we found that intercept and coefficients of `finished` and `year:finished` were extremely large. To make the results more interpretable, we subtracted `year` by 1764 (minimum year in the training data). Our model is:

`logprice` ~ `year` + `Surface` + `nfigures` + `engraved` + `prevcoll` + `paired` + `finished` + `relig` + `lands_sc` + `portrait` + `materialCat` + `year:finished` + `year:lrgfont` + `Surface:artistliving`

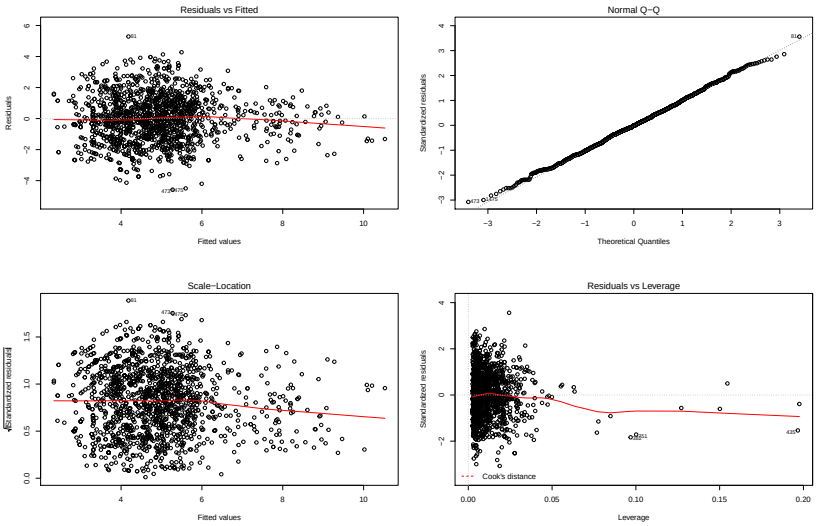
Model selection

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Without Interaction	1486	3859.135				

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
With Interaction	1483	3355.431	3	503.7036	74.2073	0

We would like to see if the interaction terms picked out by HPM is significantly increasing our prediction power here. By ANOVA, we observe that the improvement compared to model without interaction terms is statistically significant.

Residuals



Constant variability of residuals.

To determine if the model exhibits constant variability of residuals, we generate a residuals versus fits plot. In the plot, the fitted values of the model are plotted on the x axis, and the residuals of the model are plotted on the y axis. We observe that the fitted values generally form a horizontal band around the residual = 0 line, indicating overall constant variability of residuals. While we note the presence of potential outliers - observations 81, 1073, and 1475 - the plot does not exhibit a clear pattern that is indicative of non-linearity, and we are satisfied that the assumption of constant variability of residuals is met.

Nearly normal residuals.

To determine if the model has nearly normal residuals, we generate a normal probability plot. In the plot, the data are plotted by residuals generated from a theoretical normal distribution[7]. The plot for the data follows a precise linear trend, except in the extreme tail areas of the distribution. Here, we notice that the data at the extreme lower end of the data range occur at larger than expected values, and the data at the extreme upper end of the data range occur at lower than

[7] referenced from "Normal Probability Plot", available at `https://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm`

expected values (except for observation 81, which is identified as a potential outlier). This is indicative of data that is concentrated in the center of the distribution, with less data in the tails[8]. However, overall the deviation is slight, and we conclude that the model has nearly normal residuals.

[8] referenced from "A Q-Q Plot Dissection Kit", available at https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html

Homoscedasticity.

The "Scale-Location" plot is used to verify the assumption of equal variance in linear regression. If the assumption is met, the fitted values - plotted on the x axis - fall along a horizontal line with equal scatter. We observe that while the fitted values exhibit fairly equal scatter across the plot and form a general horizontal band, the trend line does exhibit a slight concave arc in the second half of the range of the fitted values, where there is less data. Again, we note that observations 81, 1073, and 1475 are identified as potential outliers. Overall, the assumption of equal variance is generally met.

Leverage and influential points.

The "Residuals vs Leverage" plot is used to determine the presence of observations with high leverage using Cook's distance. The Cook's distance values are represented by red dashed lines, and observations that fall outside of the lines are considered to be observations with high leverage. From the plot above, we observe that no observations included in the model fit fall outside of the Cook's distances, and the trend line generally follows the horizontal standardized residual = 0 line (we do note the negative trend of the line). While observations 382, 435, and 751 are highlighted as observations with potentially high leverage relative to the data, the plot does not strongly indicate the presence of any potentially influential points.

Variables

Specific summary of this model is:

	Estimate	Std..Error	2.5 %	97.5 %	Signifance
(Intercept)	3.66322	0.10679	3.45374	3.87269	***
year	0.11158	0.00880	0.09431	0.12885	***
Surface	0.00022	0.00004	0.00014	0.00030	***
nfigures	0.02791	0.00853	0.01118	0.04463	**
engraved1	0.98968	0.18189	0.63289	1.34647	***
prevcoll1	1.05053	0.18170	0.69411	1.40695	***
paired1	-0.38991	0.08489	-0.55642	-0.22341	***
finished1	1.90608	0.20115	1.51150	2.30066	***
relig1	-0.42372	0.11064	-0.64076	-0.20669	***
lands_sc1	-0.86184	0.14701	-1.15020	-0.57348	***
portrait1	-0.80113	0.21174	-1.21647	-0.38578	***

	Estimate	Std..Error	2.5 %	97.5 %	Signifance
materialCatcopper	0.38831	0.14902	0.09600	0.68063	**
materialCatother	-0.45226	0.11738	-0.68250	-0.22201	***
materialCatwood	0.24102	0.09693	0.05089	0.43116	*
year:finished1	-0.12488	0.02081	-0.16569	-0.08406	***
year:lrgfont1	0.19467	0.01438	0.16646	0.22287	***
Surface:artistliving1	0.00037	0.00011	0.00014	0.00059	**

From the summary we can see that:

1. Predictors are all sigificant.
2. There are only 3 interactions.
3. This model explains 40% of variation in training set.

Additional statistics:
Residual standard error: 1.504 on 1483 degrees of freedom
Multiple R-squared: 0.3917
Adjusted R-squared: 0.3851
F-statistic: 59.68 on 16 and 1483 DF, p-value: < 2.2e-16

This way we didn't need to do much in variable selection manually and left it to AIC and BMA. Beside this setting, we also tried start from:

1. factorizing year
2. log transforming Surface
3. both above

It turned out that do not factorize year or transform Surface lead to better results on test set.

Summary and Conclusions

Median Price for Baseline

The baseline is a painting that has these features:

1. sold at 1764
2. with ideally 0 Surface
3. has no figures
4. not engraved or finished
5. previous owner not mentioned
6. is not a pairing of other painting
7. not about regligious, landscape or portrait
8. material is canvas

9. no additional paragraph

10. artist not living

And the median price of this ideally baseline painting is $exp(3.66) = 38.86$ (livres). Confidence interval is $(31.50, 47.94)$ (livres).

Important Findings.

We notice the following:

1. If a painting is finished - that is, has a highly polished finish - then its price increases with year. If a painting is not finished, then its price decreases with year. This may be a good indicator of the quality of a painting, and thus be a good indicator for the price of a painting.

2. The effect of the surface size of a given painting varies for whether the artist of a given painting is still living, creating an interaction term.

3. While we concluded in our initial EDA that `nfigures` would likely not contribute important information for the prediction of the auction price of paintings, this variable is significant within our specified model and indicates that price increases as the number of figures in a painting increases.

Potential Limitations.

1. Our model is not appropriate for paintings that deviate greatly from the year range of 1764 to 1780. Extrapolation will occur under this circumstance, and the results of our model will not be robust.

2. Several potentially important predictors (such as `Interm`, `paired`, and `dealer`) are not included in our final specified model due to our chosen model specification process. Thus, our model may suffer from loss of predictive power if these variables are actually important for the prediction of the auction price of paintings, and this is an avenue for further analysis.

Important Interactions.

1. The estimated coefficients for `year` and `year:finished` indicates that whether or not a painting has a highly polished finish results in opposite trends of price versus year.

2. If the dealer devotes an additional paragraph (always written in a larger font size) for a given painting, price is expected to increase more with year.

3. The effect of the surface size of a given painting varies for whether the artist of a given painting is still living; price is expected to increase more with surface size if the artist of a given painting is still living.

Most important variable/interaction

Predictors	Effect on price
Year (+ 1)	**+ 19.89%**
Surface of paintings (+ 300 inches)	**+ 19.61%**
Number of figures (+ 1)	**+ 2.83%**
If artist living	**+ 0.03%**
If an additional paragraph	**+ 21.48%**
If engraved	**+ 167.98%**
If pervious owner mentioned	**+ 185.04%**
If sold as a pairing	**- 31.63%**
If finished	**+ 4.90%**
If religion related	**- 34.11%**
If a plain landscape	**- 57.17%**
If a portrait	**- 54.91%**
Material category, Copper | Canvas	**+ 47.89%**
Material category, Wood | Canvas	**+ 27.76%**
Material category, Other | Canvas	**- 35.87%**

Table of Interpretation of final model coefficients.
The first column is predictors and unit change. The second column is the effect on price. For example, others remain the same, we expect to see the price (livres) increase by 19.89% with the increase of one in sales year.

From the summary of our model, based on p-value and coefficients of variables, we identify these factors to be very important:

1. whether or not engraved or finished
2. whether or not previous owner mentioned
3. whether or not about landscape or portrait

Our Recommendations to the Art Historian.

Art is complex! Certainly, we cannot make over-generalized statements about exactly what makes a painting "more valuable", but we can provide some inference into some features that are likely to make a painting more, or less, expensive.

Less Expensive? Overall, a painting that is heavily focused on religion, landscapes, or portraiture is likely to be less expensive than a painting with other foci. Also, buyers do not seem drawn to paintings that are paired, as this on average will lead to a decrease in price.

More Expensive? Engravings done after the painting may lead to increased price, and mention of the previous owner is also likely to

increase the price of a given painting. A large painting with multiple figures depicted is likely to be expensive. The best material of support is copper - followed by wood, and then by canvas. Dealers looking to increase painting prices will include an additional paragraph with large font in the catalogue. Buyers actively looking to purchase paintings at auction should look for finished paintings with a highly polished surface.

To conclude, we find that our model indicates a myriad of various factors that may contribute to the auction price of a given painting - marketing forces, material input prices, and reknown and/or popularity of the artist are certainly all prevalent, as well as buyers' preferences of painting subject topics. We note that many of these factors are economically and societally dependent, changing dynamically from year to year. Thus, predicting the auction price of paintings, while challenging, simultaneously allows us to glimpse important factors that shape not only the sphere of art, but the spheres of economics and global culture as well.