Final Data Analysis II

Yaoyao Fan, Lucie Jacobson, Zining Ma, Jiajun Song

2019-12-15

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 categorize variables based on data type and analyze.

We first consider 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).

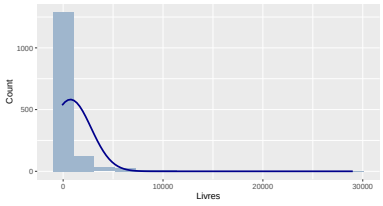


Figure 1: Histogram of Painting Price Fetched at Auction (Sales Price in Livres)

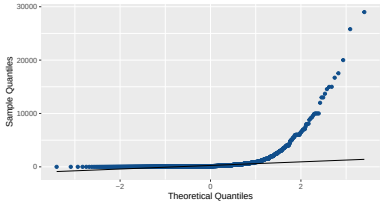


Figure 2: Normal probability plot of Painting Price Fetched at Auction (Sales Price in Livres)

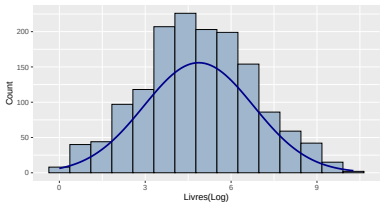


Figure 3: Histogram of Log Painting Price Fetched at Auction (Sales Price in Livres)

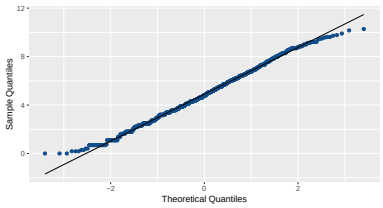


Figure 4: Normal probability plot of Log Painting Price Fetched at Auction (Sales Price in Livres)

[3] We omit variables `sale`, `subject`, `authorstandard`, `material`, `mat` at this step. Further analysis determines that these variables cause multicollinearity and interpretability issues, and furthermore do not have sufficient numbers of observations in all levels to generate robust estimates.

To 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.

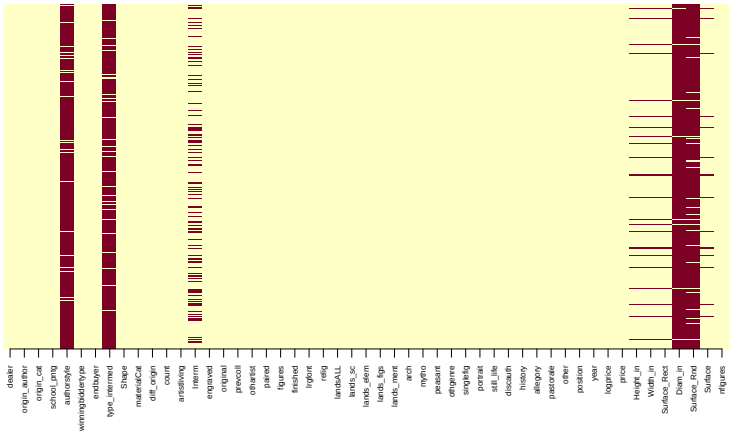
dealer	origin_author	origin_cat	school_pntg
J:201	A : 7	D/FL:594	A : 1
L:263	D/FL:590	F :483	D/FL:658
P: 93	F :578	I :170	F :608
R:943	G : 26	O :251	G : 1
	I :159	S : 2	I :193
	S : 11		S : 2
	X :129		X : 37

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 observe that data is not missing at random; the missingness is associated with our response. Thus, we cannot simply omit observations and we need to further analyze 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 specification captures information provided by the following variables:

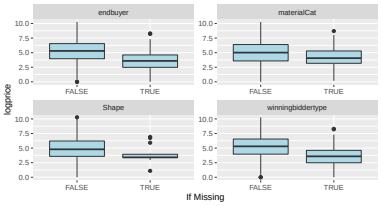


Figure 6: Missingness Effect on Response

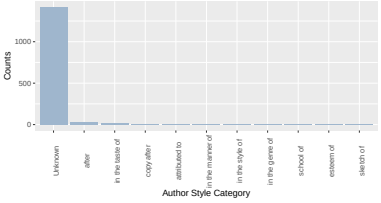


Figure 7: Counts of Author Style for Auctioned Paintings

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".

Finally, we determine that year should be a categorical variable in the data set. While time variables can be either quantitative or qualitative, it is best practice to consider year as a categorical variable: the

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).

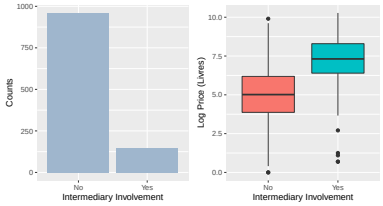


Figure 8: Painting Price and Intermediary Involvement

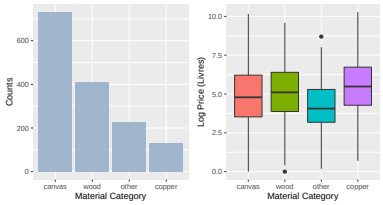


Figure 9: Painting Price and Material Category

year 1764, for example, is not an explicit measurement of 1,764 units: it is an indicator of the year of sale for a given painting. The range of `year` is (1764, 1780), which creates a factor variable with 17 levels. Given this, we opt to generate a new variable, `YearFactor`, with 6 levels:

Level 1: 1764, 1765, 1766
Level 2: 1767, 1768, 1769
Level 3: 1770, 1771, 1772
Level 4: 1773, 1774, 1775
Level 5: 1776, 1777
Level 6: 1778, 1779, 1780

This level determination, while not perfectly equal, maintains $n > 100$ observations in each level. Overall, we feel that potentially important time trends may be lost if the levels are split homogenously (resulting in year breaks), and so we opt for simple level grouping.

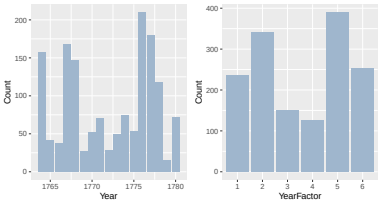
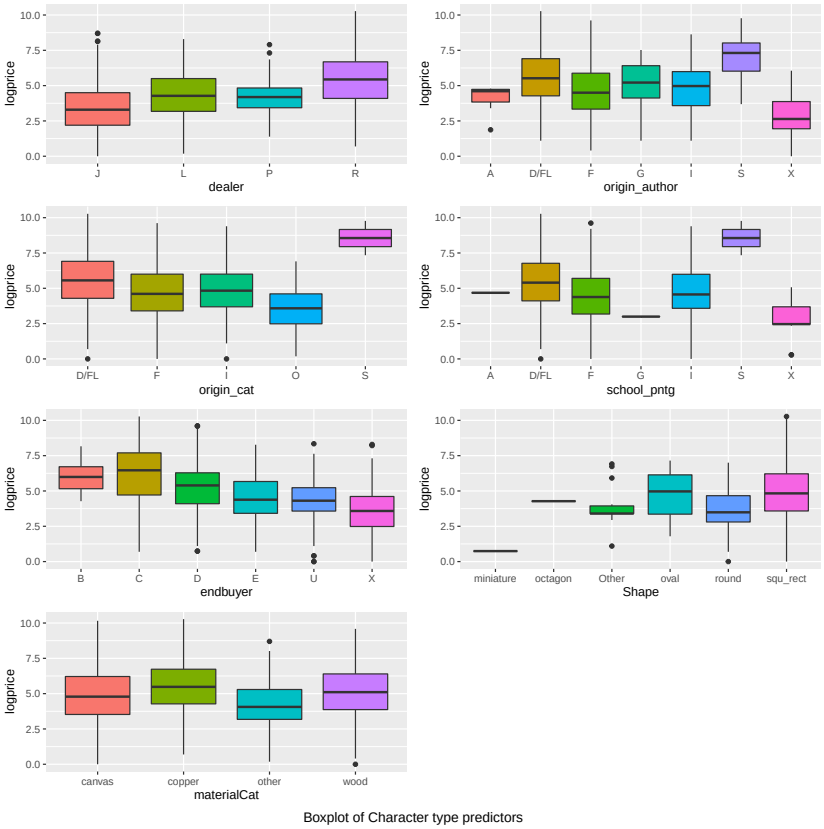


Figure 10: Transformation of Year to Group Factor

Identification of Important Variables for the Prediction of Painting Price.

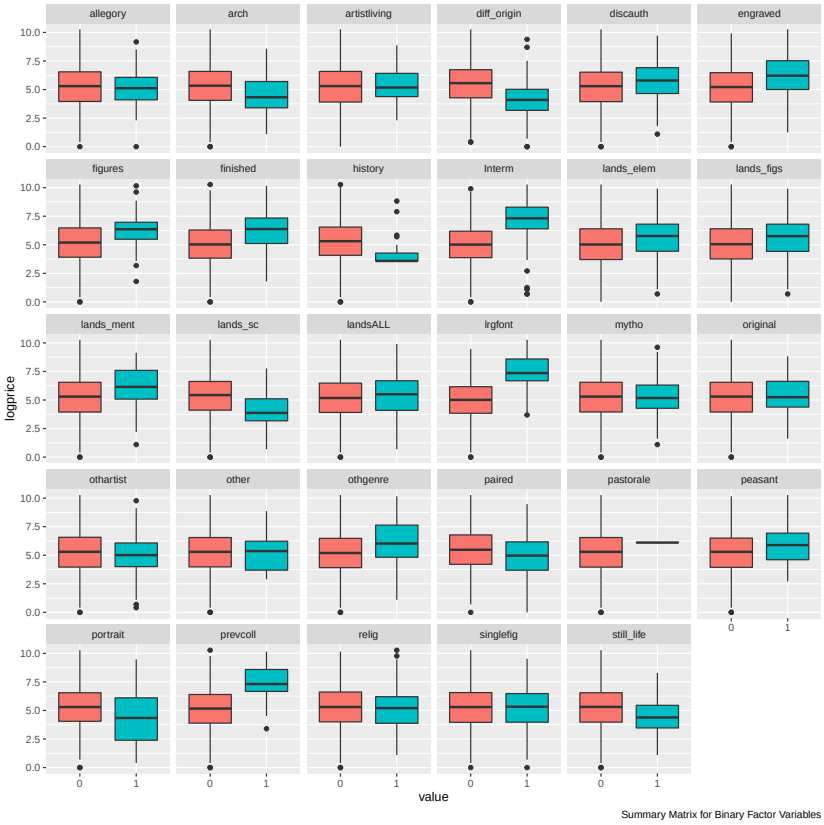
A boxplot matrix of selected variables of character type for subsequent model specification:



We note that different levels of `dealer` appear to have different

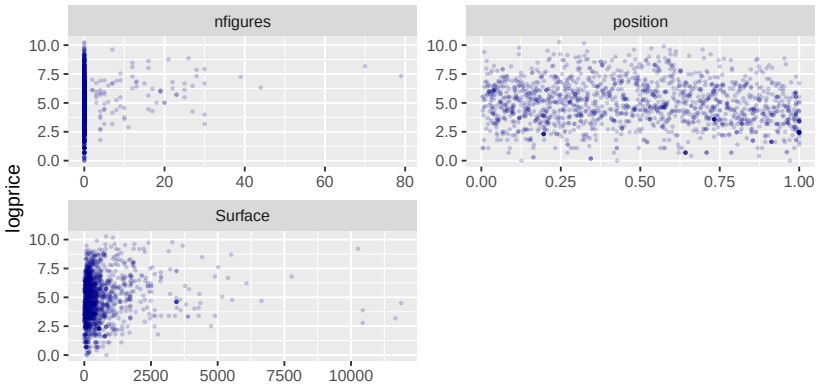
medians of sale prices, with dealer "R" maintaining a higher median sale price than other dealers. We also note that paintings with Spanish author, origin classification, and school of painting appear to have noticeably higher median sale prices than other authors, origin classifications, and schools 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 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:



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 may be a slight positive trend for paintings that do include figures. Given that this is a count variable with many zeroes, it is not appropriate to transform; previous research has shown that log-transformed count data generally performs poorly in model specification[5].

Continuing, we observe that 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 validate the transformation of `Surface`, we use the "powerTransform" method. The "powerTransform" function in R considers transformations of all variables simultaneously: both the explanatory variables and the selected response variable. This method operates under the idea that if the normality of the joint distribution of (Y, X) is improved, the normality of the conditional distribution of (Y|X) is improved. The output of the function shows the exact lambda value to which each variable should be respectively exponentiated. This makes for a quite confusing model that would be difficult to interpret. So, we consider the output values by the following rules:

- If an output value is close to 1, there is not strong evidence that a variable transformation is required.
- If an output value is close to 0.5, there is evidence that a square root transformation of the variable may be required.
- If an output is close to 0, there is evidence that a log transforma-

[5] see O'Hara and Kotze, `https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.2041-210X.2010.00021.x`

tion of the variable may be required.

Table 5: Power Transformation

	Suggest Order
logprice	0.8797628
Surface	0.0528417

From the results of the "powerTransform" method, we conclude that `logprice` does not need to be further transformed (as expected, given that this variable has already been log-transformed) and `Surface` should 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 `YearFactor`, `Surface`, `dealer`, `lrgfont`, `position`, `endbuyer`, `origin_author`, `materialCat`, `paired`, and `finished`.

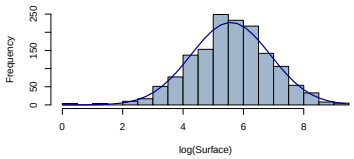


Figure 11: Surface of Painting in Squared Inches, Log Transformation

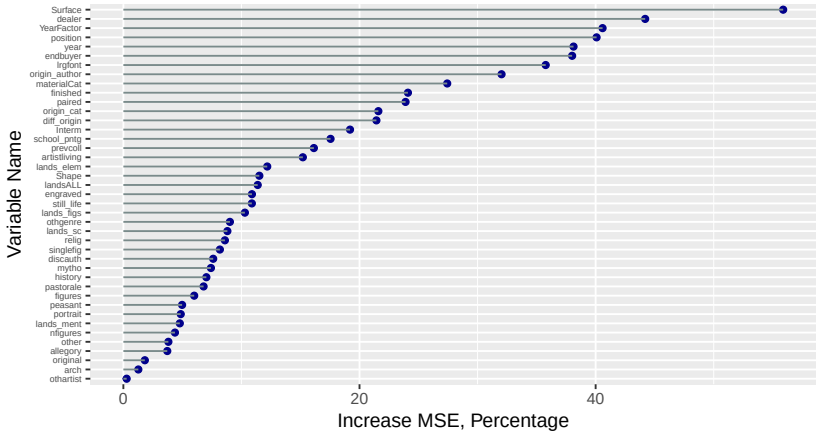


Figure 12: Variable Importance based on RandomForest

Discussion of Preliminary Model Part I.

The model we specified in Part I:

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

For specification of this model, we used Akaike information criterion (AIC) for initial variable selection. The AIC is designed to select the model that produces a probability distribution with the least vari-

ability 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. We then relied on Bayesian model averaging (BMA), which averages over models in a model class by posterior model probability to encompass the model uncertainty inherent in the variable selection problem[7], to extract the most important variables for use in our linear model. We extracted variables by obtaining the Highest Probability Model (HPM). Our resulting model explained approximately 40% of the variation in the training data (which we considered to be rather low, given the number of variables included in the model), and maintained coverage and RMSE statistics that were not better than the null model.

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

[7] referenced from "Package BMA", available at `https://cran.r-project.org/web/packages/BMA/BMA.pdf`

To improve upon our initial model, we now treat `year` as a factor variable and include `YearFactor` (please refer to EDA for a comprehensive review of this variable) in model specification instead of `year`. Furthermore, we log-transform `Surface`. Proper treatment and transformation of these variables should improve our model.

Given that `logprice` is nearly normally distributed, we do not see an immediate need to diverge from linear regression. Thus, we will again use AIC and BMA for variable selection. However, we will extract variables through the Best Predictive Model (BPM) instead of the HPM, as the BPM concludes with predictions that are closest to the Bayesian model averaging under squared error loss. Additionally, we will include more diagnostic plots to assess our model, and further analyze potential interaction terms. Then, we will consider more flexible modelling methods as needed.

Development and Assessment of Model.

With our initial modelling results and improved EDA, we decide to further explore Bayesian model averaging.

To begin modeling, we use the "bas.lm" function to conduct Bayesian adaptive sampling for Bayesian model averaging and variable selection in linear models, via sampling without replacement from a posterior distribution on models [8]. We select the Bayesian information criterion (BIC) for the prior distributions of the coefficients in the regression (approximation to the Bayes factor for large samples), and assume the model prior distribution to be the uniform distribution. Selected sampling method is Markov Chain Monte Carlo (MCMC). We choose these priors because we do not have specific information that will inform our priors, and we want to generate a model with relatively high predictive power. We specify a full model where:

[8] referenced from "bas.lm", available at `https://www.rdocumentation.org/packages/BAS/versions/1.5.3/topics/bas.lm`

- `YearFactor` is included and `year` is excluded,

- `figures` (binary) is excluded given its high association with `nfigures` (number of figures in a given painting, if specified)
- `origin_cat` and `school_pntg` are excluded to avoid multicollinearity issues with similar variable `origin_author`

```
## Warning in model.matrix.default(mt, mf,
## contrasts): non-list contrasts argument
## ignored
```

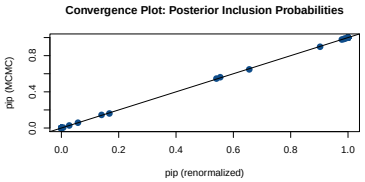
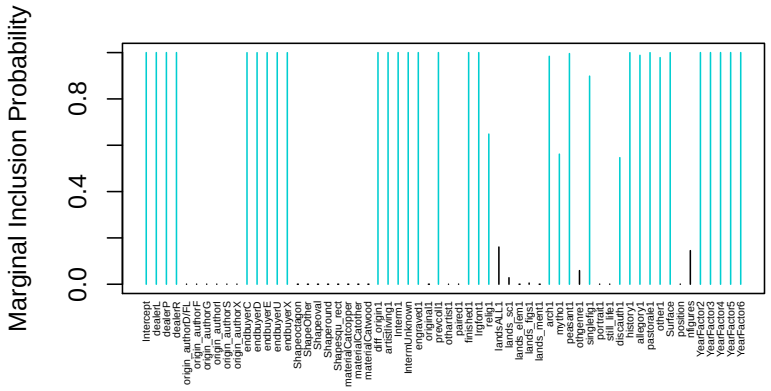


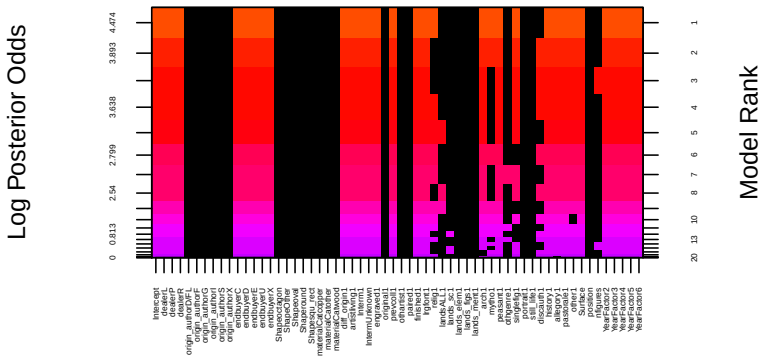
Figure 13: Covergence Plot of BMA

The plot above indicates if the posterior inclusion probability has converged under the Markov Chain Monte Carlo method. The posterior inclusion probability is the sum of all posterior probabilities associated with the models which includes a certain explanatory variable[9]. From the plot, we observe that all of the points fall on the theoretical convergence line, indicating that the number of MCMC iterations is sufficient for the data in Bayesian model averaging and do not need to be increased.

Next, we plot the marginal inclusion probability and model space:



[9] referenced from "What's the meaning of a posterior inclusion probability (PIP) in Bayesian?", available at `https://www.animalgenome.org/edu/concepts/PPI.php`



Here, explanatory variables that significantly contribute to the prediction of auction price for a given painting - that is, explanatory variables with high marginal inclusion probabilities - are highlighted in blue. From the plot, we observe that the intercept (by default), `dealer`, `origin_author`, `diff_origin`, `artistliving`, `Interm`, `engraved`, `prevcoll`, `paired`, `finished`, `lrgfont`, `lands_sc`, `portrait`, `still_life`, `Surface`, and `YearFactor` all have marginal inclusion probabilities greater than 0.5. The model space visualization provides corroboration for the previous results.

The "bas.lm" algorithm leads to a hierarchical model that represents the full posterior uncertainty after viewing the data[10]. We now want to define and generate a concrete model, namely, the best predictive model (BPM). The BPM concludes with predictions that are closest to the Bayesian model averaging under squared error loss. After generating the BPM model, we output the names of the explanatory variables included in the model. These variables are: intercept (by default), `dealer`, `origin_author`, `diff_origin`, `artistliving`, `Interm`, `engraved`, `prevcoll`, `paired`, `finished`, `lrgfont`, `lands_sc`, `portrait`, `still_life`, `other`, `Surface`, and `YearFactor`. This generally agrees with the Bayesian model averaging.

From this step, we fit a linear model with all variables identified by BPM, with additional variables identified in BMA that we feel may be important. We then use the Akaike information criterion (AIC) for further variable selection. Using this more parsimonious model, we fit a model with all possible two-way interactions to capture important interaction trends that are prevalent within the model and again use AIC to determine which variables and two-way interactions contribute significant information for the prediction of auction price of a given painting.

[10] definition referenced from "An Introduction to Bayesian Thinking: A Companion to the Statistics with R Course", available at `https://statswithr.github.io/book/stochastic-explorations-using-mcmc.html#r-demo-on-bas-package`

This results in a model that is quite overfit. Thus, we individually consider which interaction terms appear to be important. For all interactions involving levels where there are not sufficient numbers of observations, the resulting coefficient estimates are coerced to "NA". We do not include these interaction terms. Overall, the model summary indicates that the following interaction terms may be important: `dealer:difforigin`, `dealer:artistliving`, `dealer:paired`, `dealer:finished`, `materialCat:finished`, `prevcoll:finished`, `paired:lrgfont`, and `paired:YearFactor`. To briefly analyze these interactions, we generate a series of mosaic plots. A mosaic plot allows for identification of interactions between two or more categorical variables. The widths of the plot boxes correspond to the number of observations that comprise each level of the variable on the x-axis, while the heights of the plot boxes correspond to the number of observations that comprise each level of the variable on the y-axis. Overall, each plot indicates to some extent that there may potentially be an interaction effect, and we select to include all terms in subsequent model specification.

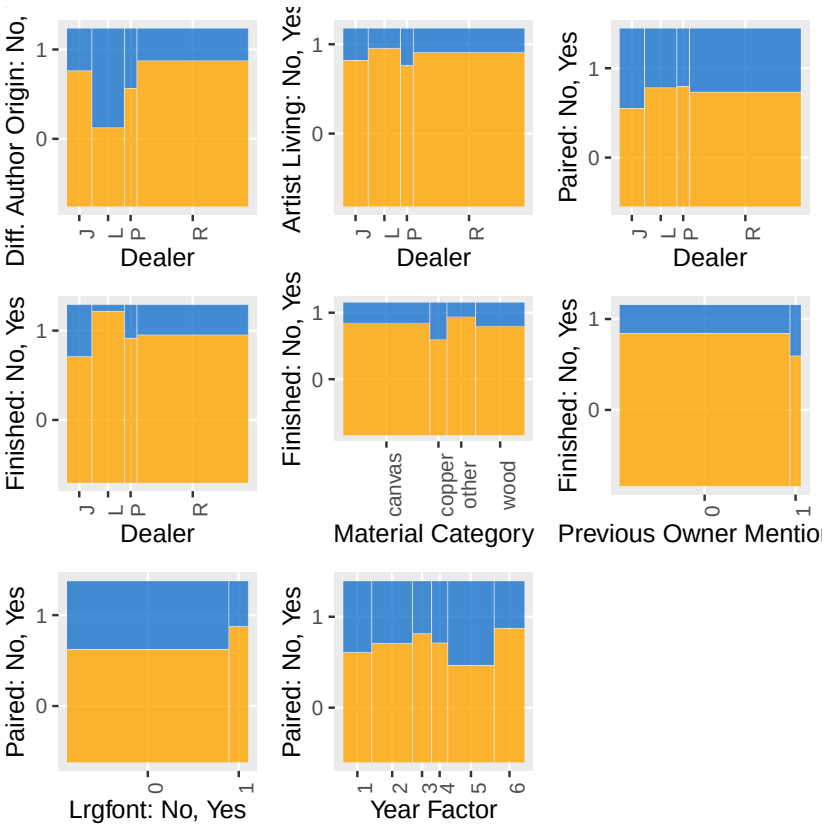
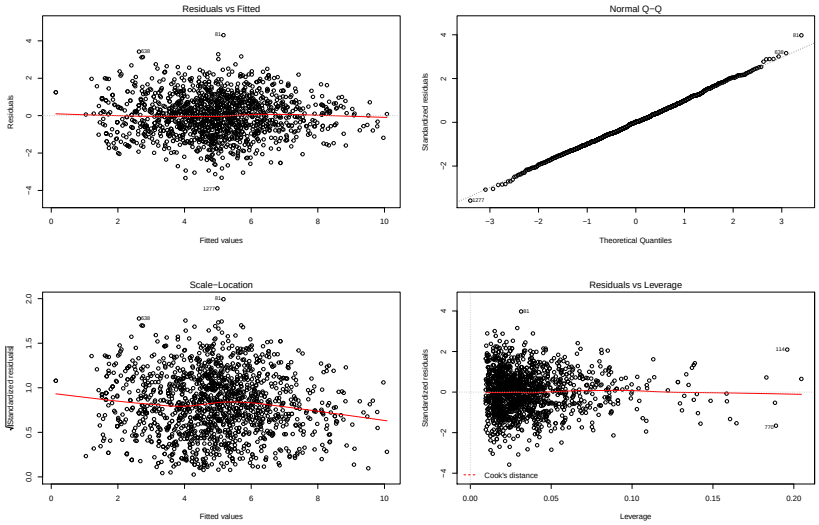


Figure 14: Mosaic plot

After fitting the model, we determine that all included variables

and terms contribute to the prediction of the auction price of a given painting. Performing an ANOVA test, the specified model is statistically significant at the $\alpha = 0.05$ level and the results indicate that the model with all eight identified interaction terms is preferred to a more parsinomious model.

Model Diagnostics.



Constant variability of residuals.

We observe that the fitted values form a horizontal line that very closely conforms to the residual = 0 line. While we note the presence of potential outliers, the plot indicates that the assumption of constant variability of residuals is met. We also note that this plot is improved in comparison to the "Residuals vs Fitted" plot for our initial model.

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[11]. The plot for the data follows a precise linear trend, and is improved from the "Normal Q-Q" plot for our initial model.

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

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. Here, we observe that the fitted values exhibit more equal scatter across the plot, forming a general horizontal band. Overall, the assumption of equal variance is 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 very closely follows the horizontal standardized residual = 0 line. While observations 81, 114, and 770 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.

Discussion of how prediction intervals obtained

For linear model, it is very convenient to get the prediction intervals for new test data, using `predict.lm(obj, newdata = testdata, interval = "prediction")`

Model testing

To test the model, we apply 5-folds cross validation on training data and see if the model has generalization error or still needs to be improved.

Table 6: Average statistics under cross validation

Bias	Coverage	maxDeviation	MeanAbsDeviation	RMSE
218.6301	0.95833	18271.60	465.9615	1424.991
218.5635	0.95200	13571.95	489.8815	1477.553

According to the summary table, first line is the evaluation metrics on training folds and second is on test folds. We observe that model achieves quite similar results on training folds or test folds, indicating that there does not exist overfitting issue. The coverage rate is satisfying. And when we continue to see how this model perform on test data, it does a rather good job actually, achieving above 95% coverage rate and around 1200 RMSE.

Variables

Specific summary of this model is:

	Estimate	Std..Error	2.5 %	97.5 %	Signifance
(Intercept)	0.47516	0.49514	-0.49612	1.44644	.
dealerL	2.59476	0.21932	2.16454	3.02499	***
dealerP	1.52684	0.25776	1.02121	2.03247	***
dealerR	2.24380	0.16101	1.92795	2.55965	***

	Estimate	Std..Error	2.5 %	97.5 %	Signifance
origin_authorD/FL	-0.04039	0.43474	-0.89318	0.81240	
origin_authorF	-0.74042	0.43656	-1.59678	0.11594	*
origin_authorG	-0.17708	0.47466	-1.10818	0.75401	
origin_authorI	-0.84589	0.44166	-1.71226	0.02048	*
origin_authorS	-0.22163	0.54902	-1.29858	0.85532	
origin_authorX	-0.89358	0.43436	-1.74563	-0.04154	*
diff_origin1	-0.12095	0.23290	-0.57780	0.33591	
artistliving1	1.13191	0.22879	0.68312	1.58070	***
Interm1	0.86578	0.10930	0.65137	1.08019	***
IntermUnknown	-0.67026	0.09249	-0.85170	-0.48882	***
materialCatcopper	0.16232	0.13490	-0.10229	0.42694	.
materialCatother	-0.24270	0.09825	-0.43544	-0.04997	*
materialCatwood	0.07205	0.08673	-0.09809	0.24218	.
engraved1	0.67280	0.13680	0.40445	0.94114	***
prevcoll1	1.12176	0.15688	0.81403	1.42950	***
paired1	0.78518	0.23970	0.31497	1.25538	**
finished1	1.07243	0.20294	0.67434	1.47051	***
lrgfont1	1.07992	0.13352	0.81801	1.34183	***
lands_sc1	-0.53886	0.11035	-0.75534	-0.32239	***
portrait1	-0.68798	0.15750	-0.99693	-0.37902	***
still_life1	-0.55755	0.15331	-0.85828	-0.25682	***
Surface	0.31739	0.02725	0.26395	0.37084	***
YearFactor2	1.08189	0.12806	0.83068	1.33310	***
YearFactor3	0.90859	0.14665	0.62092	1.19626	***
YearFactor4	1.56294	0.16605	1.23722	1.88866	***
YearFactor5	1.97140	0.13236	1.71176	2.23104	***
YearFactor6	0.98568	0.16653	0.65901	1.31236	***
dealerL:diff_origin1	-0.37768	0.27943	-0.92580	0.17045	.
dealerP:diff_origin1	-0.41213	0.32818	-1.05589	0.23163	.
dealerR:diff_origin1	-0.57504	0.23151	-1.02916	-0.12092	*
dealerL:artistliving1	-0.86159	0.32770	-1.50441	-0.21877	**
dealerP:artistliving1	-0.30106	0.37644	-1.03948	0.43736	.
dealerR:artistliving1	-0.67403	0.24905	-1.16257	-0.18548	**
dealerL:paired1	-1.26006	0.24479	-1.74024	-0.77988	***
dealerP:paired1	-1.11654	0.34319	-1.78973	-0.44334	**
dealerR:paired1	-0.22626	0.18963	-0.59824	0.14571	.
dealerL:finished1	-0.51248	0.45306	-1.40120	0.37623	.
dealerP:finished1	-1.16842	0.38448	-1.92261	-0.41423	**
dealerR:finished1	-0.52497	0.21972	-0.95598	-0.09396	*
materialCatcopper:finished1	0.34515	0.24763	-0.14060	0.83090	.
materialCatother:finished1	1.05214	0.28840	0.48640	1.61787	***
materialCatwood:finished1	0.32999	0.18749	-0.03780	0.69778	*

	Estimate	Std..Error	2.5 %	97.5 %	Signifance
prevcoll1:finished1	-1.07733	0.29902	-1.66390	-0.49077	***
paired1:lrgfont1	-0.67621	0.23149	-1.13031	-0.22211	**
paired1:YearFactor2	-0.98709	0.20731	-1.39375	-0.58042	***
paired1:YearFactor3	-0.84221	0.26284	-1.35780	-0.32662	**
paired1:YearFactor4	-0.94210	0.28068	-1.49269	-0.39152	***
paired1:YearFactor5	-0.83563	0.20330	-1.23443	-0.43684	***
paired1:YearFactor6	0.35972	0.24819	-0.12714	0.84658	.

Additional statistics:
Residual standard error: 1.137 on 1447 degrees of freedom
Multiple R-squared: 0.6608
Adjusted R-squared: 0.6486
F-statistic: 54.2 on 52 and 1447 DF, p-value: $<$ 2.2e-16

From the summary table of variable estimates and confidence intervals, we find out that almost all the predictors are statistically signifcant in terms of 0.05 level.

With interaction terms included and increase of number of factor levels, it does not make much sense to talk about the interpretation of one single predictor as it is closed related to other predictors indicated by the model. We still could observe important variable or variable combinations that make a painting expensive. For example, pictures with large font introduced in the subject is expected to be 219.17% more expensive than those not. Also, pictures with type R dealer and not living artist is expected to be 14.67% more expensive than those with other type dealer and artist still live.

To pursue a model with lowest RMSE on test set, we have developed another model which we think is also very interesting and meaningful to included in the report.

Random Forest & Linear Regression Model.

Now the question is: can we improve the model?

To improve the performance of our model, we now consider a two-part model: first, we introduce the tree-based method of random forests to fit an approximate value, and then fit the residual using linear regression. This implementation is similar to boosting, where trees are grown sequentially, using information from previously fit trees.

Overall, the approach is intuitive: we first classify a given painting in a large class, and then account for differences based on painting features.

Here, we note that for the variable `nfigures`, most observations are 0 with the remaining observations sparsely distributed over a large range (please refer to EDA for graph). Thus, we will log-transform this count variable for subsequent model specification.

After log-transformations, we find that there are linear relationships for both log(Surface) and log(nfigures) on range where the value is greater than 0. Our objective is to fit a linear specification for observations with values greater than 0, and fit a point estimation for observations with values equal to 0.

We use a such a method: (Take log-Surface as an example)

1. Create an indicator 'logS_n0', which equals 1 if Surface does not equal 0, and equals 0 otherwise.

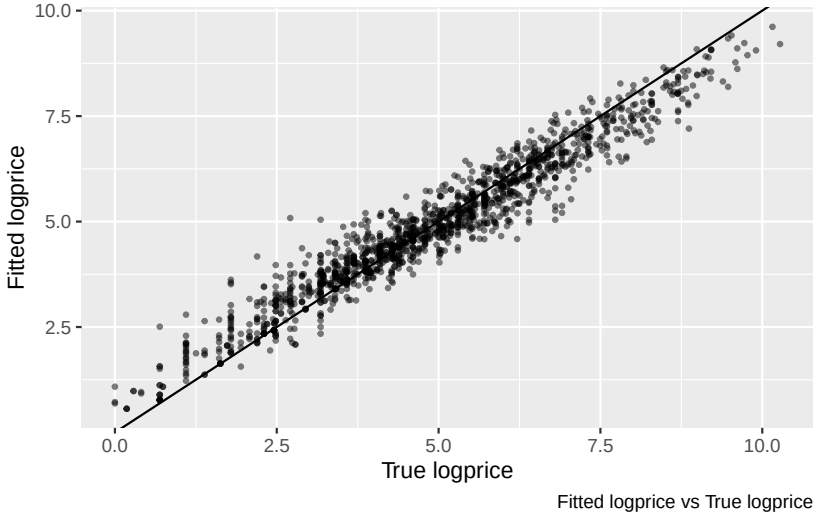
2. Use 'logS_n0 + logS' in the model formula. Thus, the final model could be:

$$fit \mid (Surface > 0) = b_1 + k * log\_S$$
$$fit \mid (Surface = 0) = b_0$$

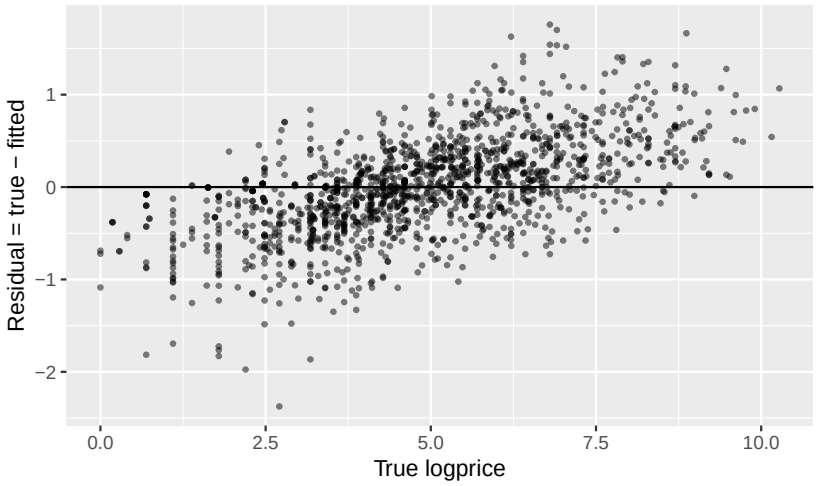
where $b_0 \neq b_1$.

Random Forest.

First, we fit a random forest using relevant predictors (selected from previous linear modelling) and analyze fit:



In this plot, we find that when the true value is low (<5), the model tends to overestimate. When the true value is high (>5), it will underestimate. Thus, it is clear that there is a pattern between residuals and `logprice`. Analyzing further, we have:



Residual vs True logprice

There is a clear linear relationship between the residuals of the random forest model and the true values. So, we consider fitting the residuals using linear regression in the next step.

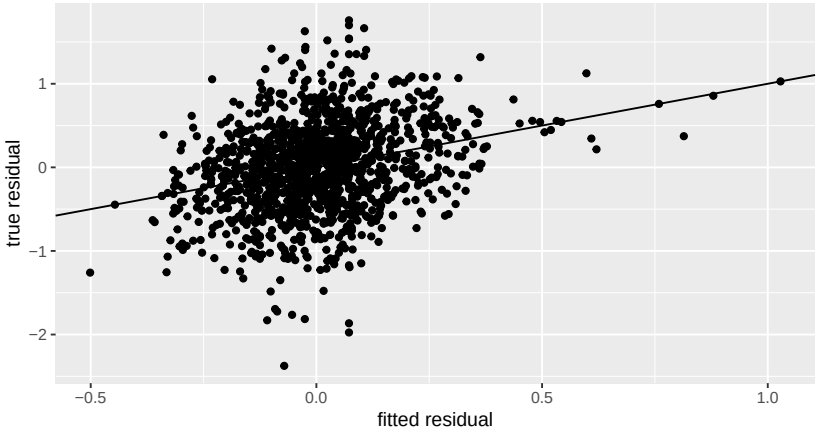
Linear Regression for Residual.

We now fit a linear model with the predictor variables and significant interactions identified in previous modelling efforts.

Interactions included here are:

- `dealer:diff_origin`
- `engraved:prevcoll`
- `prevcoll:finished`
- `paired:lrgfont`
- `paired:year`
- `materialCat:finished`

How well we fitted residual:



Prediction Interval

Our final prediction is of the form:

$$\hat{y} = \hat{y_{rf}} + \hat{residual_{rf}}$$

Our assumption of the model is:

$$y = \hat{y_{rf}} + \hat{residual_{rf}} + \epsilon$$

We can estimate $Var(\epsilon)$ by

$$\hat{Var(\epsilon)} = Var(y - \hat{y})$$

And we can obtain prediction interval of $\hat{residual_{rf}}$ from the linear model.

Unfortunately, we cannot obtain a prediction interval of $\hat{y_{rf}}$. Instead, what we could do is to give an under-estimated prediction interval based on $Var(\epsilon)$ and $\hat{residual_{rf}}$.

Sepcifically, we assume that:

$$y \sim Normal(\hat{y}, Var(\epsilon) + Var(\hat{residual_{rf}}))$$

Thus the prediction interval is:

$$\hat{y} \pm 1.96 * \sqrt{Var(\epsilon) + Var(\hat{residual_{rf}})}$$

Model evaluation

Random Forest

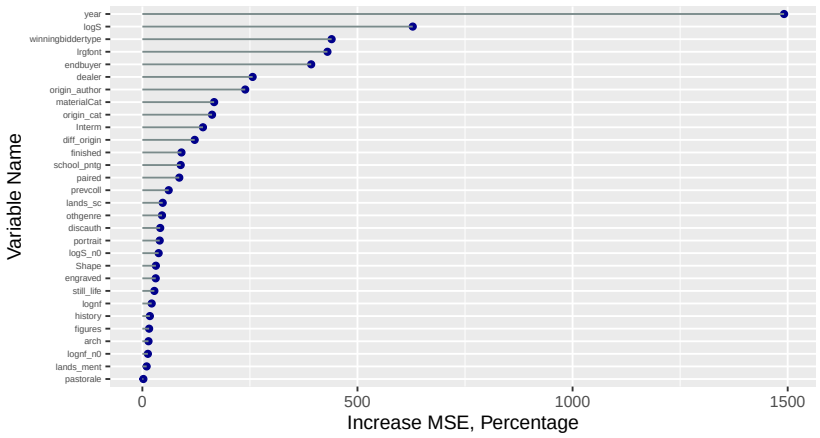


Figure 15: Variable Importance based on RandomForest

From importance plot of random forest model, we see that year, Surface, winning bidder type, large font, end buyer, dealer and original author are important in decision making. Disparate to what we

observed in EDA, 0/1 factor predictors and nfigures do not appear to play a relatively important role in the prediction of auction price for a given painting.

Linear Regression for Residual.

Summary table:

R	R^2	R^2_{adj}	σ^2
0.29	0.08	0.02	0.51

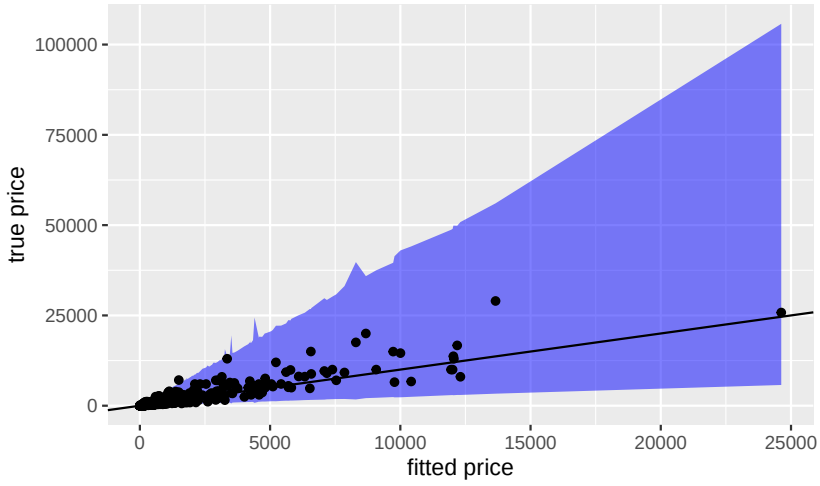
Objectively, we should conclude that this linear regression on residual is not significant. There are only 5 significant variables at the $\alpha = 0.05$ level. Furthermore, the model does not explain more than 10% of total variation.

However, we include the regression in our model because there is a distinct difference in performance on the test data: with this linear regression step, RMSE on test data could be under 1000; RMSE could be greater than 1200 if we omit this step.

Besides the discussion of whether to include or omit this step in our model, we also tried variable selection with AIC. Although this does result in increased significant predictors, performance on test data is worse (with RMSE greater than 1200). Thus, we decide to utilize the full linear model in this step.

Performance

Coverage on training data (in livre):



Coverage plot on training data

We can see that the prediction interval covers most of the true values, and overall provides a decent fit on points with large values.

From the residual plot and qq-plot, we see that:

1. Residuals are not distributed equally on the fitted data range. Variance of residual tends to be smaller at the two ends and becomes greater in the middle.
2. The model tends to overfit when the fitted value is less than 5, and tends to underfit when the fitted value is greater than 5.
3. The assumption on our model $y = \hat{y_{rf}} + \hat{residual}_{rf} + \epsilon$ is not correct, because the residuals are not normally distributed, and the residual depends on true value or fitted value.

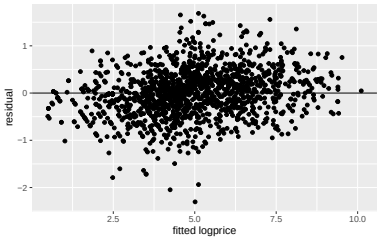


Figure 16: Residual plot on training data

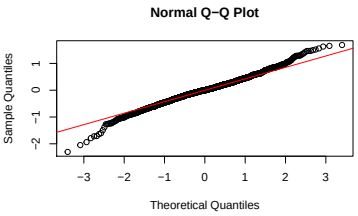


Figure 17: QQ plot for residuals

Model Evaluation

Before we have access to test data, we use k-fold cross validation on training data to estimate model performance:

Our estimation of Bias, Coverage, maxDeviation, MeanAbsDeviation, RMSE are:

Table 9: Evaluation on train set

Bias	Coverage	maxDeviation	MeanAbsDeviation	RMSE
257.97	0.86	13787.5	456.68	1437.18

Our score on test set is:

Table 10: Evaluation on test set

Bias	Coverage	maxDeviation	MeanAbsDeviation	RMSE
144.39	0.91	8046.72	336.36	912.85

Compared to other groups, we have advantages on all of these scores except coverage. We have lowest Bias, maxDeviation, MeanAbsDeviation and RMSE.

Model result

Our prediction of paintings in validation data tells us that these paintings may have highest prices:

lot	author	predicted price
28	Pierre Paul Rubens	10071.169
170	Philippe Wouwermans	9470.833
108	Nicolas Berghem	8899.874
159	Adrien Vanden Veld	7441.713
30	Gerard Dow	6917.518
7	Barthelemi Etienne Murillos	6416.164
167	Karel du Jardin	6268.131
51	Rembrandt Van Rhyn	5909.465
171	Eustache Le Sueur	5699.962
118	Isaac Van Ostade	5387.827

Considering important variables we identified in development process, we observed that these paintings have common features in these variables:

lot	year	winningbiddertype	lrgfont	endbuyer	dealer	origin_author
28	1777	C	1	C	R	D/FL
170	1767	DC	1	C	R	D/FL
108	1777	C	1	C	R	D/FL
159	1776	DC	1	C	R	D/FL
30	1769	EBC	1	C	R	D/FL
7	1769	EC	1	C	R	S
167	1776	DC	1	C	R	D/FL
51	1777	DC	1	C	R	D/FL
171	1777	DC	1	C	R	F
118	1777	DB	1	B	R	D/FL

Model Disadvantages.

1. Our model does not allow for the calculation of a precise prediction interval. As we mentioned in discussion of residual plot, our model is based on an assumption that is not validated by our data. Furthermore, we cannot estimate a prediction interval for a tree model, and thus we are unable to define a theoretical α-level of our prediction interval.

2. It is challenging to interpret our model and fully explain the effects of variables on painting prices because 1. we use random forest, and 2. we incorporate many predictors in the linear model for fitting residuals.

3. The performance of our model on test set is likely not representive. Our estimation of model performance based on k-fold validation is worse than our score on test data. There is a distinct likelihood that our model will not perform equally well on validation data.

What We Could Do Better.

If we had additional time to work on this project, we would ideally first focus more on data cleaning and EDA. We think it would be valuable to explore all of the painters in the data set, and determine potential associations between painter characteristics and auction price. Additionally, it would be interesting to undertake text analysis on the `subject` variable, which contains a short description of subject matter. This could be accomplished utilizing the "sentimentr" package, where an average sentiment score for a text vector can be generated. By converting text to a numerical value, we could analyze if the subject of a painting is associated with other explanatory variables:

- for instance, are different types of buyers more likely to purchase paintings with negative sentiment scores (negative emotionally charged painting subjects), positive sentiment scores (positive emotionally charged painting subjects), or neutral sentiment scores (scores approaching zero)?

and, of course, if the subject of a painting provides significant information for the prediction of auction price.

Furthermore, although we do feel that linear regression is appropriate for the log-transformed price, we would like to further explore variable transformations and more flexible models.