STA 521 Final Project Part II

Team 10: Bin Han, Jingyi Zhang, Jonathan Klus

12 December 2018

1. Introduction

In this study, the auction prices of paintings in 18th century Paris were examined. Specifically, we wish to understand the variables which affect the prices of the paintings, and then be able to predict auction prices based on characteristics of a certain painting. By fitting an appropriate model, we will also be creating a tool to help decide whether specific paintings that are either underpriced or overpriced given their realization of the covariates that were included in the model.

One of the main challenges in building this model is to narrow down the number of covariates from the 59 candidates in the original data set to less than 20 in the final model. This must be done in such a way that an undue amount of bias is not introduced, and overfitting is avoided. Another challenge is to properly deal with the messiness of the data, including both missingness, covariates with a very large number of levels, multicollinearity in the data, and discrepancies in data entries (e.g. same category marked differently).

The ability to explain the results and provide recommendations to individuals without statistical background is equally important and challenging, since the primary audience for this analysis is intended to be art historians. The goal was therefore to balance predictive performance, model simplicity, and interpretability in order to create a pricing model for artwork in 18th century France.

2. Exploratory data analysis

A) Data summary & cleaning

To begin, we looked at the summary of the original training data. There are few numeric variables and a lot of binary variables. Some variables, such as `Interm`, `Surface`, `Height_in` etc. have mising values, which needed to be imputed. The following steps were taken to clean the data:

a. The first step was to reduce the dimensionality of the problem by removing variables that were deemed not to be useful due to their being summarized more succinctly by another similar variable, having too many levels, or not containing any useful information (i.e. taking on the same value for each observation). These variables included: `lot`, `sale`, `price`, `count`, `subject`, `authorstandard`, `winningbidder`, and `other`. From the summary table, the `count` variable has all 1's; the `other` variable does not convey useful information; the other variables, such as `names` and `subjects`, are not useful in predicting the response variable (such as names). From the table of unique values we can see that some categorical variables have over 1,000 unique values. Therefore, we chose to remove them in the first step. The alternative to this would be to attempt to recode the variable in an effort to preserve some of its information for the model. In Part I, the `author` variable receive this treatment, but in the second iteration of this model, we chose to recode based on the perceived value of the names of several top artists. Only the authors with more than 10 paintings are kept as a distinct level, and all others were coded as `other`.

b. By further screening the variables, we found out that `Surface` and `Surface_Rnd`, `Surface_Rect` are highly correlated, as they are all measurements based on the value of `Height_in`, `Width_in`, and `Diam_in`. We decided to use `Surface` in our initial model since it contained the most information about all of these measurements. The same issue happened to `material`, `mat`, and `materialCat`. `materialCat` recodes the other two more succinctly, therefore, we used `materialCat` for ease of modeling and interpretation. We applied the same strategy to keep `landsALL` and get rid of other variables related with landscape.

c. This data contained a great deal of structurally missing values (i.e. missingness resulting from how the researchers coded the data, rather than truly unavailable or omitted information). For those variables that have multiple levels, to be consistent with how the data was originally coded, we recoded the missing levels as “X”, which stands for either “other” or “no information” in the code book, depending upon the variable in question. For `materialCat` and `Shape`, since there are so many levels, we grouped some levels with few (<10) observations together, coded as the “other” group. The remaining binary vairables were converted into factors.

d. The remaining data issue was how to deal with missing values in the numeric continuous variable `Surface` and the binary variable `Interm`. The `mice` package (Multivariate Imputation by Chained Equations) was used to address this problem. It uses the observed values of other covariates in the dataset to create a model to impute the missing values. This method is superior to complete case analysis, which would result in losing an unacceptably large amount of data, as well as simpler imputation methods (i.e. imputing the mean of a given covariate to replace missing values).

e. The variable `subject` also contained many levels, potentially with many observations representing the same realization but expressed differently (i.e. varied spelling and capitalization). Strings of the same meaning were detected by observation and recatogorized. From the resulting values, levels with more than 20 observations were kept while all others were coded as `other`.

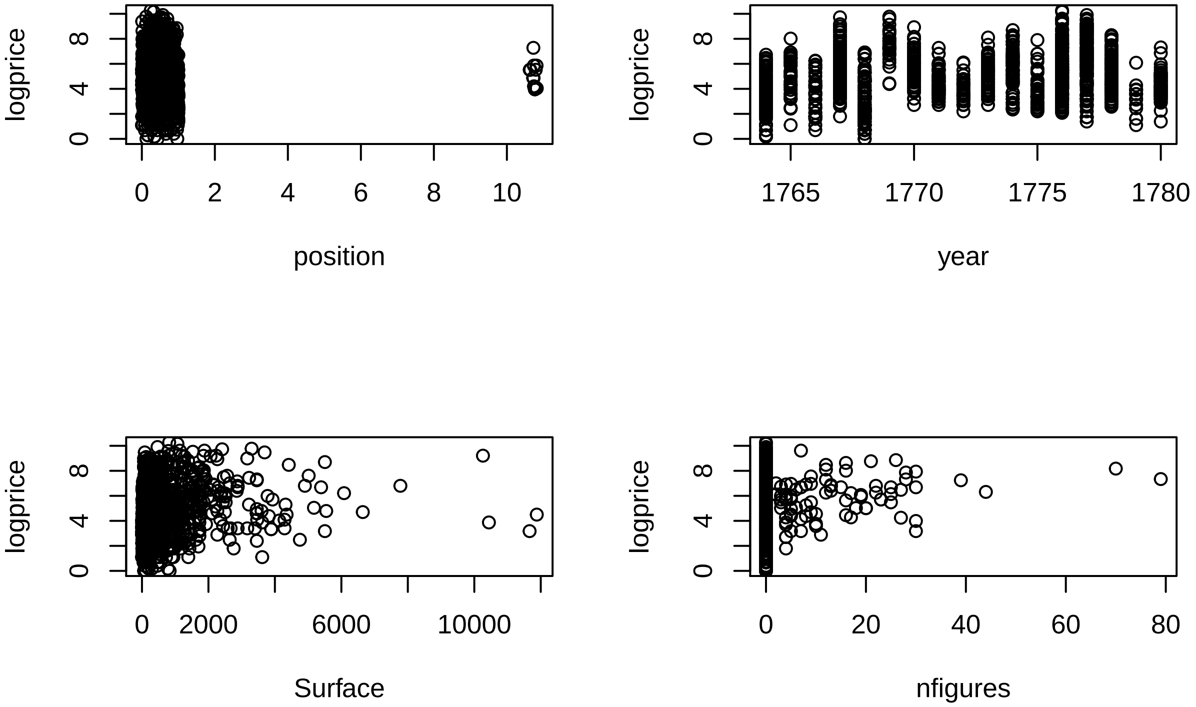
author	sum
other	1307
David Teniers	46
Philippe Wouvermans	27
Francois Boucher	26
Charles de la Fosse	17
French	16
Gasparo Van Vitelle	13
Rosalba Carriera	13
Gaspard Netscher	12
Nicolas Poussin	12
Nicolas Berghem	11

subject	sum
other	605
Paysage	324
People	194
Saint	121
Portrait	43
Fruit$Flower	42
Adoration	39
Arch	31
Buste	30
Marine	25
Battle	23
Sujet	23

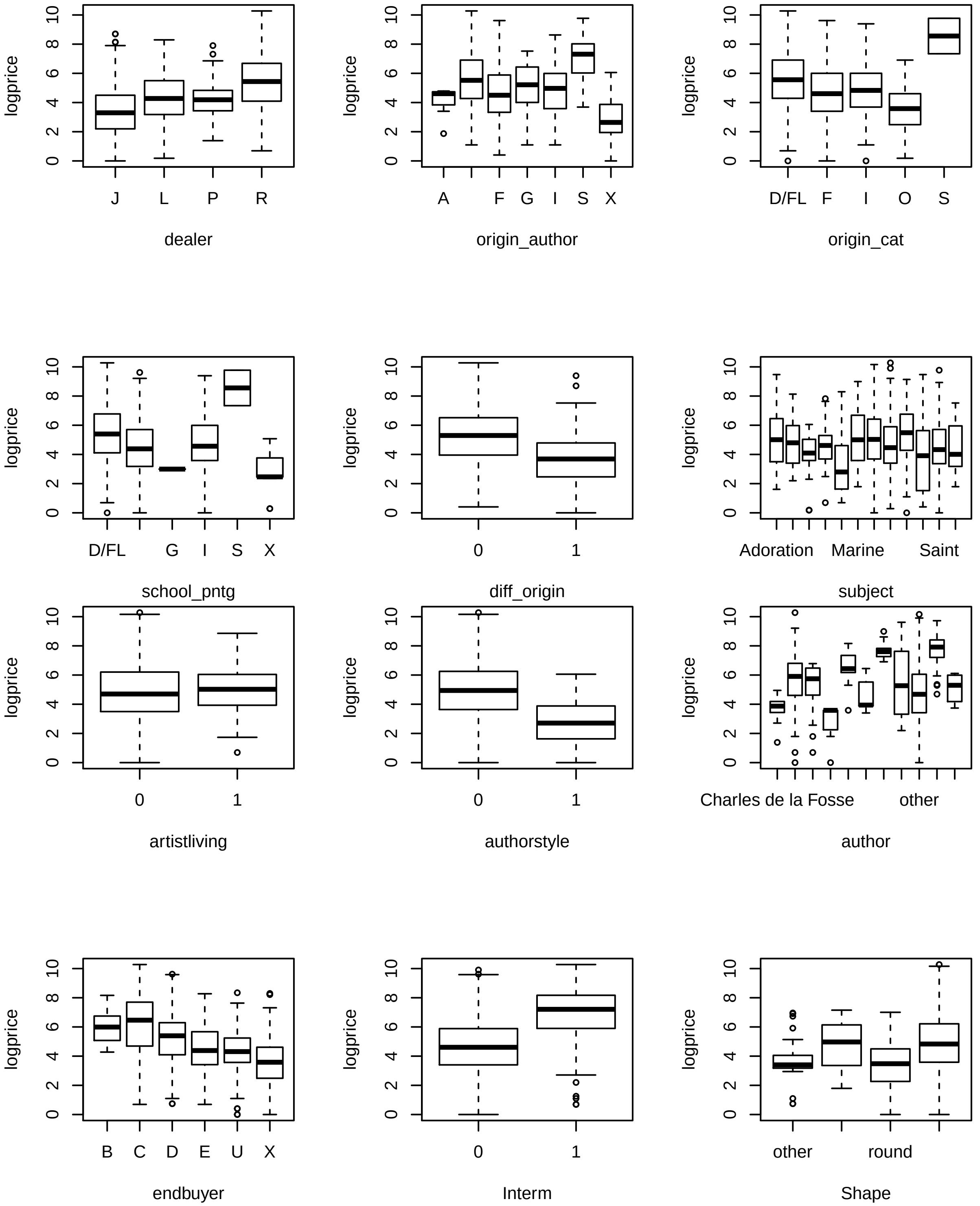
B). Plots

The relationship between the remaining features and the response `logprice` was then further examined . Using scatter plots, we can roughly determine which variables should be put into the initial model by visual inspection if they appear to have a linear relationship with logprice. For categorical variables, we want to check if the logprice spans different ranges in different levels by plotting them using boxplots. For numeric variables, we want to check if there is a clear relationship between them and logprice.

For numeric variables, we see that Surface and nfigures seem to show a weak but positive relationship with logprice. Since there are several extremely large values in position (potentially outliers), it is hard to see that real pattern between the majority of points and logprice. These variables will be kept in the initial model, but may potentially be removed later in the development process.



Since there are 33 categorical variables, we don't show the boxplots for all of them. But we have applied the same method to check all the categorical variables. The following variables show some differences in logprice at different levels (not yet considering the magnitude of the difference at this time): subject, author, dealer, origin_author, origin_cat, school_pntg, diff_origin, authorstyle, endbuyer, Interm, Shape, materialCat, engraved, prevcoll, figures, finished, Irgfont, othgenre, discauth, and still_life.



If we were to choose best variables for prediction at this stage, we would consider the magnitude of differences in `logprice` in the different levels of each categorical variable, and the strength of relationships. The 10 variables chosen were: `Surface`, `subject`, `author`, `dealer`, `school_pntg`, `diff_origin`, `authorstyle`, `endbuyer`, `Interm`, `prevcoll`, `engraved`, `Irgfont`.

For numeric variables, we note that `Surface` and `nfigures` appear to have a weak but positive relationship with `logprice`. Since there are several extremely large values in `position` (potential outliers), it is difficult to know if there is a truly useful relationship here between the majority of points and `logprice`.

3. Discussion of preliminary model

The overall characteristics of the model that we built in part I were: relatively low bias, reasonable coverage (~95%), and high RMSE compared to other teams. The methodology used to arrive at the first model included initial EDA, followed by BMA and stepwise selection with AIC in order to narrow the number of potential covariates to include in the model by seeking predictors with the highest posterior inclusion probabilitys (BMA) and highest information content (AIC/BIC). These method do not perform an exhaustive search for all possible models, thus the true model and the best model for prediction might not have been captured. This is likely due to the fact that interaction terms were not included in the BMA step due to the computational intensity of such a calculation. Furthermore, a few important variables were likely excluded from the initial model out of hand (e.g. `author`, `subject`), and thus a good deal of important information was likely lost. These covariates were recoded and will be included in the model selection process during this second phase.

Since there is inherently a tradeoff between bias and RMSE in any modeling problem, it is reasonable that we were able to achieve relatively low bias, while RMSE was relatively higher. Both metrics may be improved with a better model, which may end up being something other than a linear model, or through deeper data cleaning. In the second phase, additional attention will be focused on tree/forest methods, as well as further development of the linear model to determine which provides superior prediction for the problem at hand.

4. Development of the final model

Summary of Covariates Included and their Coefficients

	Coefficient	2.5%	97.5%
(Intercept)	-246.114	-526.795	34.566
Shapeoval	0.455	-0.287	1.197
Shaperound	-0.747	-1.434	-0.061
Shapesqu_rect	0.255	-0.270	0.779
school_pntgF	-0.537	-0.698	-0.375
school_pntgG	-2.362	-4.601	-0.123
school_pntgI	-0.540	-0.753	-0.326
school_pntgS	0.498	-1.121	2.117
school_pntgX	-0.922	-1.341	-0.503
dealerL	1.771	1.381	2.162
dealerP	0.520	0.069	0.972
dealerR	1.824	1.524	2.124
Interm1	-0.411	-1.186	0.365
paired1	1.111	-0.953	3.175
artistliving1	94.414	15.817	173.010
diff_origin1	0.173	-0.284	0.629
endbuyerC	5.733	-278.256	289.722
endbuyerD	108.564	-174.555	391.683
endbuyerE	-78.867	-366.507	208.774
endbuyerU	63.997	-222.666	350.660
endbuyerX	99.265	-183.453	381.983
finished1	0.692	0.494	0.891
year	0.141	-0.018	0.299
Surface	0.001	-0.001	0.002
portrait1	-0.461	-1.026	0.103
still_life1	-0.301	-0.748	0.145

	Coefficient	2.5%	97.5%
prevcoll1	1.631	0.777	2.486
authorstyle1	-0.895	-1.201	-0.590
lrgfont1	1.251	0.944	1.557
subjectArch	-0.589	-1.415	0.238
subjectBattle	-0.042	-0.939	0.856
subjectBuste	-0.623	-1.280	0.033
subjectFruit$Flower	-0.164	-0.897	0.570
subjectMarine	0.012	-0.717	0.740
subjectother	-0.028	-0.479	0.424
subjectPaysage	-0.333	-0.809	0.143
subjectPeople	0.196	-0.287	0.680
subjectPortrait	-0.275	-1.061	0.511
subjectSaint	-0.173	-0.666	0.320
subjectSujet	0.156	-0.630	0.942
discauth1	0.364	-0.030	0.759
authorDavid Teniers	0.762	-0.006	1.529
authorFrancois Boucher	0.359	-0.459	1.176
authorFrench	-0.347	-2.136	1.442
authorGaspard Netscher	0.878	-0.086	1.841
authorGasparo Van Vitelle	1.247	-0.217	2.711
authorNicolas Berghem	1.790	0.597	2.983
authorNicolas Poussin	1.527	0.604	2.451
authorother	0.492	-0.151	1.135
authorPhilippe Wouvermans	1.299	0.430	2.168
authorRosalba Carriera	0.502	-0.713	1.717
dealerL:Interm1	1.245	0.161	2.330
dealerP:Interm1	1.124	-1.349	3.596
dealerR:Interm1	1.309	0.503	2.116
dealerL:paired1	-0.778	-1.295	-0.261
dealerP:paired1	-0.416	-1.069	0.237
dealerR:paired1	-0.114	-0.522	0.293
dealerL:artistliving1	-0.754	-1.434	-0.074
dealerP:artistliving1	-0.651	-1.486	0.185
dealerR:artistliving1	-0.707	-1.336	-0.079
dealerL:diff_origin1	0.065	-0.500	0.630
dealerP:diff_origin1	-0.542	-1.222	0.138
dealerR:diff_origin1	-0.672	-1.152	-0.192
artistliving1:endbuyerC	1.451	-0.295	3.197
artistliving1:endbuyerD	1.655	-0.074	3.383
artistliving1:endbuyerE	2.036	0.194	3.878
artistliving1:endbuyerU	1.674	-0.100	3.447
artistliving1:endbuyerX	1.554	-0.211	3.318
artistliving1:finished1	-0.300	-0.835	0.235
artistliving1:year	-0.054	-0.098	-0.009
diff_origin1:Surface	0.000	0.000	0.000
diff_origin1:portrait1	0.530	-0.334	1.395
diff_origin1:still_life1	-1.145	-1.818	-0.472
diff_origin1:prevcoll1	0.086	-0.770	0.942
endbuyerC:Surface	-0.001	-0.002	0.001
endbuyerD:Surface	-0.001	-0.002	0.001
endbuyerE:Surface	0.000	-0.002	0.001
endbuyerU:Surface	0.000	-0.002	0.001

	Coefficient	2.5%	97.5%
endbuyerX:Surface	-0.001	-0.002	0.001
paired1:endbuyerC	-1.351	-2.763	0.061
paired1:endbuyerD	-0.985	-2.389	0.419
paired1:endbuyerE	-0.608	-2.080	0.865
paired1:endbuyerU	-0.849	-2.293	0.595
paired1:endbuyerX	-0.997	-2.432	0.438
endbuyerC:year	-0.003	-0.164	0.157
endbuyerD:year	-0.061	-0.221	0.098
endbuyerE:year	0.044	-0.118	0.207
endbuyerU:year	-0.037	-0.199	0.125
endbuyerX:year	-0.057	-0.216	0.103
portrait1:authorstyle1	0.094	-1.777	1.965
Interm1:lrgfont1	-0.266	-0.749	0.217
paired1:lrgfont1	-0.588	-1.072	-0.104
paired1:subjectArch	0.496	-0.680	1.673
paired1:subjectBattle	-0.744	-2.014	0.527
paired1:subjectBuste	1.462	0.220	2.704
paired1:subjectFruit$Flower	0.176	-0.936	1.287
paired1:subjectMarine	-0.721	-1.943	0.501
paired1:subjectother	-0.177	-0.996	0.642
paired1:subjectPaysage	0.105	-0.731	0.940
paired1:subjectPeople	-0.156	-1.033	0.721
paired1:subjectPortrait	0.025	-1.053	1.102
paired1:subjectSaint	0.331	-0.613	1.276
paired1:subjectSujet	-0.351	-1.584	0.881
paired1:discauth1	-0.344	-0.975	0.286
paired1:authorDavid Teniers	-0.946	-2.409	0.517
paired1:authorFrancois Boucher	0.983	-0.651	2.617
paired1:authorFrench	0.534	-1.699	2.767
paired1:authorGaspard Netscher	0.532	-1.646	2.709
paired1:authorGasparo Van Vitelle	-0.509	-2.464	1.447
paired1:authorNicolas Berghem	0.138	-1.751	2.028
paired1:authorNicolas Poussin	0.091	-2.657	2.839
paired1:authorother	-0.083	-1.381	1.215
paired1:authorPhilippe Wouvermans	0.674	-0.899	2.247
paired1:authorRosalba Carriera	0.075	-1.759	1.909
finished1:discauth1	0.768	0.239	1.298
lrgfont1:discauth1	-0.753	-1.444	-0.063
dealerL:prevcoll1	-0.622	-1.857	0.612
dealerP:prevcoll1	-1.755	-3.366	-0.144
dealerR:prevcoll1	-0.690	-1.604	0.224
Interm1:prevcoll1	-0.402	-1.016	0.211

Variables:

a. The base variables that were chosen for the final model include: `Shape`, `school_pntg`, `dealer`, `Interm`, `paired`, `artistliving`, `diff_origin`, `endbuyer`, `finished`, `year`, `Surface`, `portrait`, `still_life`, `prevcoll`, `authorstyle`, `lrgfont`, `discauth`, `subject` and `author`. These were chosen based upon their inclusion in the Best Predictive Model found using Bayesian Model Averaging (via the BAS package).

b. The interactions chosen include: `dealer*Interm`, `dealer*paired`, `dealer*artistliving`, dealer*diff_origin, artistliving*endbuyer, artistliving*finished, artistliving*year, diff_origin*Surface, diff_origin*portrait, diff_origin*still_life, diff_origin*prevcoll endbuyer*Surface, endbuyer*paired, endbuyer*year, authorstyle*portrait, Interm*lrgfont, paired*lrgfont, paired*subject, paired*discauth, paired*author, finished*discauth, lrgfont*discauth, prevcoll*dealer, and prevcoll*Interm. Best AIC and BIC selection were used as a tool in determining which interactions should be included in the final model. The models which minimized information loss (i.e. with the highest AIC and BIC) were used as a guide to select interactions for inclusion in the final model.

c. Partial Explanations:

- dealer: the type of dealer that the auction went through significantly affects the price of the painting. For example, compared with dealer J, the average price from dealer L is `179% higher`. (Same interpretation for dealer P and R, with different coefficients)
- finished: if the painting is noted for being highly finished, the selling price on average is `69.76% higher` than when the painting is not noted for being highly finished.
- prevcoll: when the previous owner is mentioned, the average selling price is `128.0% higher` than when the previous owner is not mentioned.
- lrgfont: when the dealer devotes an additional paragraph, the average selling price is `124.9% higher` than when there is no additional paragraph.
- authorstyle: when the author's name is introduced, the average selling price is expected to be `88.39% lower` than when the author's name is not introduced.
- author: which author painted the painting also has some influence on the price. Compared with author Charles de la Fosse, author David Teniers' paintings are `80.57% higher` in price on average. Author Nicolas Berghem's paintings are `181.4% higher` in price on average.
- dealer&Interm interaction: when an intermediary is present, which the price of the auctioned paintings differs significantly among different dealers. For instance, if the dealer is R and an intermediary is used, the average selling price is `107.0% higher` than when the dealer is J with an intermediary.
- finished*discauth: given that the painting is noted for being highly finished, when the dealer engages with authenticity, the average price is expected to be `76.2% higher`.
- diff_origin:still_life: given that the origin of painting based on nationality of artist is different from the origin of painting based on dealer's classification, if the description indicates still life elements, the price is expected to be `-112.8% lower`.

Variable selection/shrinkage:

In developing the final model, several different methods were explored to induce

a. Linear Model

The linear model from part I does a fairly good job in prediction, though it could clearly be improved. Therefore, after cleaning two more variables from the original dataset, we decided to refine the linear model first. The plan was to add new features and interactions into the model in order to explain more variation in the response variable and improve out-of-sample RMSE. Similar as the process in part I, we applied BMA (Bayesian Model Averaging) to select the base variables that have high posterior probabilities and include them in the initial model. Then we tested all possible interactions and used AIC to select interactions that are good for prediction. However, the output from AIC contained too many interactions, and using all of them would have likely led to the problem of overfitting the training data. Additionally, it contains some interactions with coefficients that are not estimable (i.e. result in NA output), and some that do not make sense at all. Such results were manually removed, and the resulting features were combined to find the best final linear model in terms of the highest adjusted R-squared and lowest in and out-of-sample RMSE.

b. Tree Model

Since this data set contains many categorical variables, a tree-based model would seem appropriate to address these interactions when modeling logprice. We tested two tree models: random forest and boosting (bagging is not feasible in this case as we have 39 variables, which is beyond the limit of possible selection candidates at each node). As for random forest, we used $mtry = 13$ as this is a regression problem. For boosting, we used 5000 trees and tried different interaction depth (4, 6, 8, 10). Both methods perform better than the linear model from part I for in and out-of-sample prediction, but do not perform as well as the linear model generated above. While they may achieve similar or only slightly higher out-of-sample RMSE on the test, their bias is nearly double that of the linear model.

c. Poisson & Negative Binomial Regression

Even though not strictly count data, the auction price might be be treated in such way so that we can use poisson & negative binomial regression. We trained the tested the poisson model first. With all the variables included (even with all the interactions), the in-sample residual deviance is 100 times higher than the residual degrees of freedom. We concluded that the poisson model was not appropriate and proceeded with negative binomial model. The in-sample deviance was still inferior to that of the linear model is part I (even with all the interactions) and the problem of over-dispersion remained.

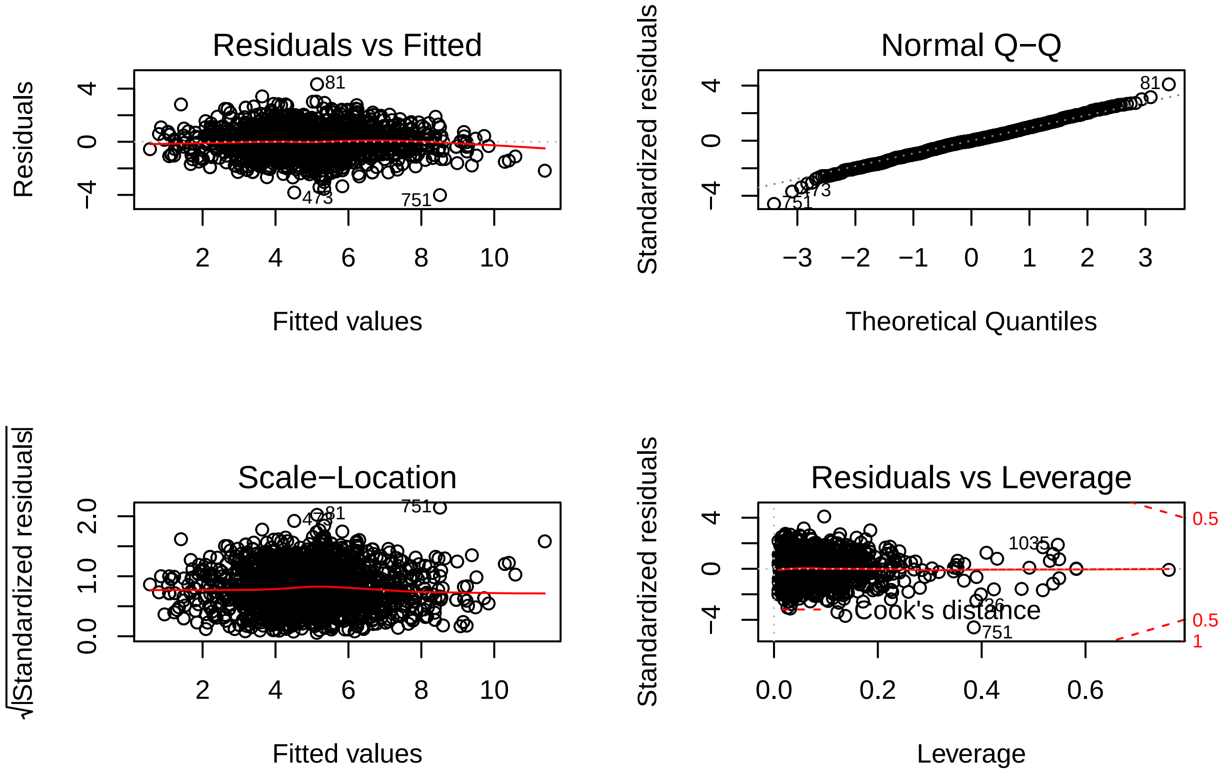
d. Xgboost

Similarly to in section (b) above, as a result of the large number of potential interactions and categorical variables present in this model, a tree-based approach seemed like it might be a good approach. The xgboost package (Extreme Gradient Boost) has garnered a lot of attention for its good predictive performance, so this was implemented and we attempted to tune the model by alterating the allowed depth of trees (tried 4, 5, and 6), as well as the regularization penalties. Ultimately, this method proved to be nearly as accurate as the linear model, but with significantly higher bias. It was therefore not selected as the final model for this problem.

Comparing all the models that were fitted above, it was concluded that the linear model had the best performance in terms of prediction (lowest RMSE both in and out-of-sample in almost all cases), and unbiasedness (lowest bias out of all fitted models by nearly a factor of 2). The linear model was relatively more complex than the linear model fit in part I, resuting in some loss of interpretability (e.g. via interactions that aid in prediction, but are difficult to meaningfully interpret). However, it is still more interpretable than tree-based models, particularly because it includes predictors both with and without interactions. Therefore, we concluded that the best model was the linear model.

Residuals

The final model that is specified has generally good in-sample performance and characteristics. There are three outliers in the residuals versus fitted values plot, though the plot otherwise appears to indicate that the model meets the homoskedasticity assumption. The residuals also appear to meet the normality assumption, based upon a visual inspection of the Normal Q-Q plot. Though there is some slight deviation in the tails, there does not appear to be any extreme variability or overall pattern to the residuals that would indicate an underlying distribution other than the normal.



Prediction Intervals for Out-of-Sample Data

Table 4: Prediction Interval

fit	lwr	upr
100.312	11.028	912.464
286.842	20.851	3946.011
174.185	18.346	1653.792
4753.640	411.347	54934.407
450.037	47.955	4223.384
285.933	27.753	2945.920
133.708	14.719	1214.646
4.985	0.496	50.071
111.329	11.842	1046.618
117.379	11.945	1153.431
30.053	3.156	286.219
452.450	49.327	4150.106
439.650	46.638	4144.528
1018.663	110.584	9383.577
17.368	1.894	159.219
90.828	10.031	822.431
2168.869	236.275	19908.954
1067.533	116.055	9819.742
103.452	9.434	1134.437
16.001	1.609	159.112

Since we are still using the linear regression model, we used `predict` function with `interval = "pred"` argument to obtain the prediction interval. In the table above, the prediction intervals for the first 20 out-of-sample observations are displayed for the sake of space.

5. Assessment of the final model

Model evaluation

The final model appears to perform well relative to the model fit from part I and to the alternative methods that were detailed above. The final out-of-sample diagnostics were an RMSE value at 1210.42 and bias of 184.08. Coverage was also just above 95%, which is acceptable given that we wish to evaluate the model at a 95% confidence level.

The scale-location plot does indicate three residuals that are right at the border of being large (standardized residuals greater than or near a value of 2.0). For case 751, this is likely due to its above-average surface area of 11,880 sq. in., but its relatively low price of 90 livres (compared to an average of approximately 130 livres in the training data). A similar argument may be made for case 473. For case 81, the opposite seems to be the problem. The painting is relatively small, with a surface area of just 84.4 sq. in., but sold for approximately 13,000 livres. In these few cases, there is likely some characteristic of the painting that is not being fully captured by the predictors, though this does not appear to have an overall great affect on the performance of the model. Of concern might be that if this model were used to value a painting with similar characteristics, we may undervalue it.

The residuals vs. leverage plot makes note of three cases that are potentially influential points: 423, 1129, 1351. They have a Cook's distance of one, indicating potentially high influence. The removal of these variables was considered, but a refit of the model revealed that there was not an overall large effect of these three observations on the greater model. Furthermore, eliminating them led to additional points being flagged as points of high leverage and potential influence. This exemplifies the potential for a race to the bottom as more extreme observations are removed, resulting in other observations appearing more extreme when compared to the remaining data. Overall, we did not believe that there was enough reason to warrant eliminating any observations from the training data.

The model was further evaluated using added variable plots to understand each predictor's contribution to the model, and whether any transformations may be worth exploring. However, this process was not very helpful because the model includes so many categoricals variables, and so few numeric continuous variables. There was no indication from visual inspection of the avPlots that transformations of the numeric predictors would have yielded improved model fit.

Finally, the magnitude of coefficients relative to their standard errors was considered when evaluating the model. This relationship is well-documented by the value of the t-test statistic. While many coefficients were not found to have p-values that fell below our alpha-level of 0.05, these values may still aid in prediction. Furthermore, many correspond to the different levels of a particular categorical variable. While some levels of the variable may not have been found to have a coefficient that is statistically significantly different than 0 at the 95% level of confidence, other levels of the same variable may have such a result. So the variable is overall important to the model, and eliminating it or recoding its levels post-hoc would eliminate information and may even introduce additional and unwarranted bias to the model.

Model testing

After selecting the final model and predicting with the test dataset, we evaluated several metrics to evaluate the performance of the model. The two most important evaluations of the model when comparing it to other candidates were its out-of-sample RMSE value, which represents the predictive performance of the model, and its bias, which measures the tendency for the model to systematically over- or underestimate (i.e. the model the captures the true data-generating process would have zero bias).

In terms of the final model that we fit, we observed out-of-sample RMSE of 1210.42 and bias at 184.08. Since the method that we end up using is a linear model, this is superior to our initial linear model and on par with results obtained using other methods discussed in section 4. The use of stepwise AIC as the model selection criteria has the goal result of a model that has better predictive performance (compared to BIC, which seeks the “true model”). Coverage was also calculated and found to be just above 95%, which was a slight improvement over the initial model and on par with our expectations for how an acceptable model would perform. Of note is that, given a different test set, it is possible that these results could be alterated. This is one potential downfall of the train-test set setup that was utilized here, and a method like k-fold cross validation with a witheld validation set would likely yield less variable results in this regard.

Lastly, since we aim at balancing the simplicity, interpretability and performance of the model, certain levels of categorical variables were recoded and not all interactions were included in the model, even though the stepwise AIC process may have suggested their inclusion. This could also negatively impact the RMSE and bias of the final model, but would also add complexity and the potential for overfitting the training data.

Model result

price	rownum
2736.9	998
1067.9	983
1058.6	987
798.9	982
576.9	994
573.1	992
538.6	981
443.0	988
351.2	996
291.3	984

According to our model, the most valuable ten paintings in the validation set are, from most expensive to the tenth expensive, marked with row number 2065, 2066, 2543, 1071, 2538, 2584, 2517, 2477, 2022 and 2528. All of these paintings tend to share similar characteristics like their shape, dealer, school, endbuyer, whether they are paired or not, whether they are engraved or not and etc. Most notably, five of the top ten paintings have the same artist (author), Berghem. Unsurprisingly, in the model, all else equal, a painting by Berghem is associated with a price increase of 179% relative to artist la Fosse, who does not appear in the top ten here. Most of the other shared features of these 10 paintings, after taking into account the related interaction terms, had a positive association with the price. Therefore, the model ended up predicting high prices for these paintings.

6. Conclusion

	Adj_R2	RMSE	Coverage
Part I Model	0.6129	1502.99	0.9506667
Part II Model	0.6634	1219.52	0.9587000

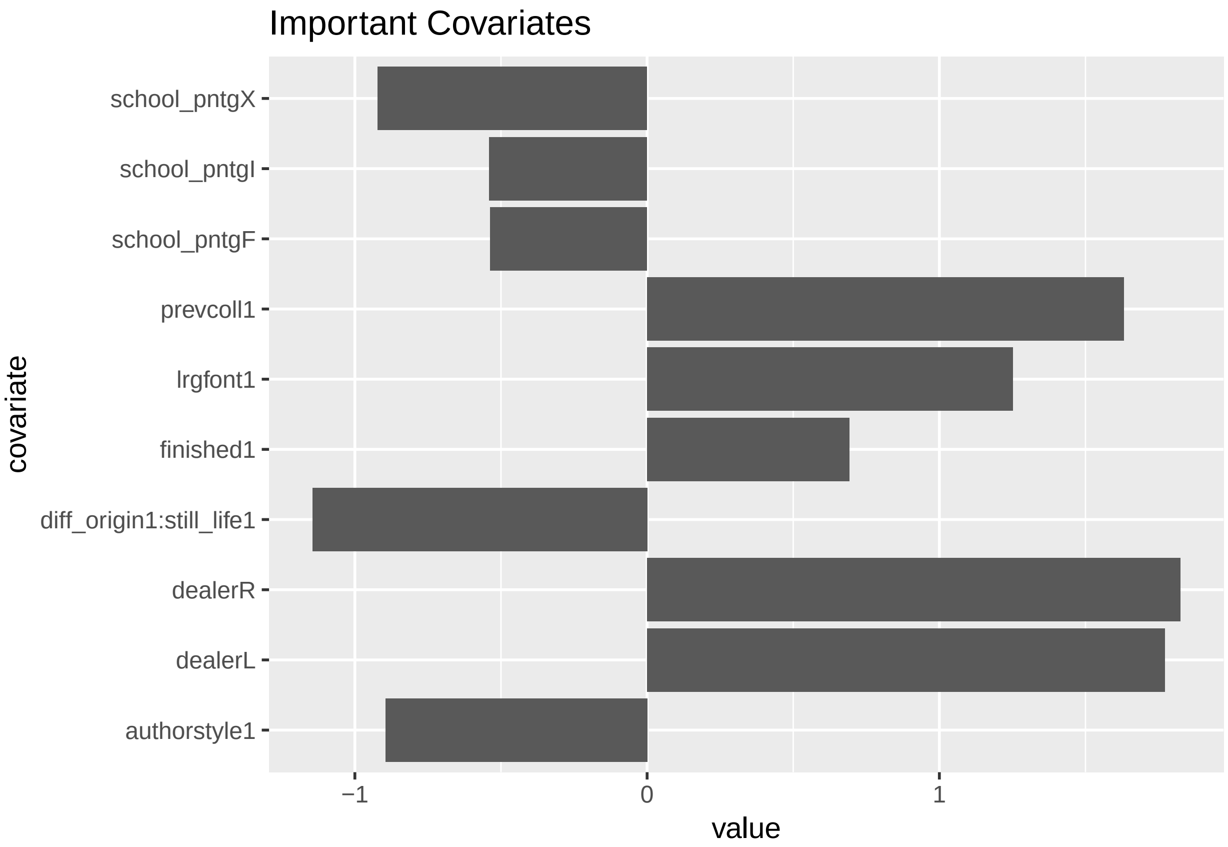
Even though we ended up using linear models for both parts I and II, we do notice that in sample adjusted R-squared, RMSE and coverage, have all improved from the preliminary model.

The changes made to reach this result included deeper data cleaning, more careful attention paid to potential outliers, as well as reasonable manual adjustments to predictor selection. Through the data exploration and cleaning, model selection, and testing process, several important lessons have been learned: - In realistic data analysis problems, we rarely get access to nicely formatted, clean data without any missingness. Therefore, when encountered such data sets, a deep and careful data cleaning process is important before moving on to further modeling and testing steps. In terms of properly treating missing data, simply deleting the observations with missing data might cause lost of valuable information, thus approriate data imputation is also important. Once modified to a cleaner, complete dataset, the later modeling results would turn out to be more precise, less computationally costly, with the maximum amount of available information for model training. - Ideally, we are able to find a model that is both the closest to the true model and has the best prediction performance. However, in this case, we had to choose to balance interpretability with predictive performance in seeking the best model for our intended audience. - More advanced models like tree/forest methods, xgboost and etcetera, aren't always preferable to ordinary linear models. - Even though the auto selection methods provided candidate models selected based on certain metrics (e.g AIC, BIC), not all predictors and interactions should always be included for reasons of interpretability, simplicity, and potential for overfitting. The ultimate purpose of the studies is not only to fit the best model, but also to draw conclusions and make inferences for a real-life scenario, and potentially serving other studies, whose researchers don't necessarily have statistical background.

As far as the questions of interest of this particular study is concerned, we've learned that the auction price of a painting is not solely dependent on the immediately observable qualities of the painting itself. It also heavily depends on other factors involved in the auction process. For example, which types of endbuyers are intrigued by the painting, or which dealer the painting is auctioned from, can also significantly impact the final auction price.

```
ols.2.ci =  confint(ols.2)
coef_table = data.frame(names = rownames(ols.2.ci),
                        coef  = coef(ols.2),
                        lower = ols.2.ci[,1],
                        upper = ols.2.ci[,2],
                        p_val = summary(ols.2)$coefficients[,4])[-1,] %>%
            filter(p_val < 0.05) %>%
            arrange(p_val)

ggplot2::ggplot(data = coef_table[1:10,], mapping = ggplot2::aes(x = names, y = coef)) +
  ggplot2::geom_bar(stat = "identity") +
  coord_flip() +
  xlab("covariate") +
  ylab("value") +
  ggtitle("Important Covariates")
```



Even though most paintings with certain characteristics like large surface area end up with an auction price on the higher end, some paintings with exceptionally large area can be valued a lot less than expected, which resulted in a couple of noted outliers in this scenario. Therefore, it is crucial for historians to consider all features of a painting and decide how much each factor should be weighted when predicting the price. In particular, our analysis has found that characteristics like the dealer involved in the transaction may have large influence on the price at which a painting is sold, as can the amount of space that the dealer devotes to the description of the painting (i.e. the lrgfont variable), if the painting has a pedigree (i.e. if the previous owner is mentioned), and if it is considered highly finished by experts. These variables are summarized in the chart above. There are several other important variables, and the above summary is meant only to serve as a snapshot of a few of the most influential for an art scholar who was interested in the subject.