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 due to

their, upon visual inspection, appearing 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 predictive variables for predicting, we would consider the magnitude of differences and the strength of relationships. The 10 variables we chose are: `Surface`, `subjetc`, `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, 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 and BIC 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 exhausive 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)	-253.208	-535.780	29.363
Shapeoval	0.453	-0.288	1.193
Shaperound	-0.755	-1.440	-0.071
Shapesqu_rect	0.257	-0.266	0.780
school_pntgF	-0.532	-0.694	-0.371
school_pntgG	-2.361	-4.598	-0.123
school_pntgI	-0.532	-0.745	-0.318

	Coefficient	2.5%	97.5%
school_pntgS	0.485	-1.132	2.103
school_pntgX	-0.950	-1.369	-0.532
dealerL	1.803	1.413	2.193
dealerP	0.526	0.075	0.978
dealerR	1.819	1.519	2.118
Interm1	-0.167	-0.977	0.644
paired1	1.168	-0.906	3.243
artistliving1	91.811	13.287	170.335
diff_origin1	0.201	-0.254	0.656
endbuyerC	11.389	-274.458	297.236
endbuyerD	117.091	-167.953	402.134
endbuyerE	-70.584	-360.143	218.976
endbuyerU	73.995	-214.661	362.651
endbuyerX	104.240	-180.397	388.878
finished1	0.693	0.495	0.892
year	0.145	-0.015	0.304
Surface	0.001	-0.001	0.003
portrait1	-0.470	-1.034	0.093
still_life1	-0.315	-0.762	0.131
prevcoll1	1.541	0.704	2.378
authorstyle1	-0.887	-1.192	-0.581
lrgfont1	1.248	0.942	1.554
subjectArch	-0.680	-1.509	0.149
subjectBattle	-0.066	-0.963	0.830
subjectBuste	-0.653	-1.307	0.002
subjectFruit$Flower	-0.182	-0.915	0.551
subjectMarine	-0.027	-0.754	0.700
subjectother	-0.060	-0.511	0.391
subjectPaysage	-0.355	-0.830	0.119
subjectPeople	0.165	-0.317	0.648
subjectPortrait	-0.321	-1.106	0.464
subjectSaint	-0.212	-0.703	0.280
subjectSujet	0.116	-0.669	0.901
discauth1	0.374	-0.021	0.768
authorDavid Teniers	0.796	0.026	1.565
authorFrancois Boucher	0.412	-0.407	1.231
authorFrench	-0.351	-2.140	1.437
authorGaspard Netscher	0.942	-0.018	1.901
authorGasparo Van Vitelle	1.292	-0.171	2.756
authorNicolas Berghem	1.815	0.622	3.008
authorNicolas Poussin	1.511	0.590	2.432
authorother	0.522	-0.122	1.166
authorPhilippe Wouvermans	1.336	0.466	2.206
authorRosalba Carriera	0.532	-0.684	1.747
dealerL:Interm1	0.736	-0.347	1.819
dealerP:Interm1	-0.543	-2.386	1.299
dealerR:Interm1	1.099	0.260	1.938
dealerL:paired1	-0.708	-1.226	-0.189
dealerP:paired1	-0.432	-1.085	0.222
dealerR:paired1	-0.085	-0.493	0.323
dealerL:artistliving1	-0.825	-1.504	-0.147
dealerP:artistliving1	-0.599	-1.426	0.228

	Coefficient	2.5%	97.5%
dealerR:artistliving1	-0.726	-1.355	-0.098
dealerL:diff_origin1	0.002	-0.561	0.566
dealerP:diff_origin1	-0.491	-1.174	0.193
dealerR:diff_origin1	-0.686	-1.166	-0.207
artistliving1:endbuyerC	1.529	-0.214	3.272
artistliving1:endbuyerD	1.691	-0.035	3.418
artistliving1:endbuyerE	2.062	0.223	3.902
artistliving1:endbuyerU	1.710	-0.062	3.481
artistliving1:endbuyerX	1.613	-0.150	3.376
artistliving1:finished1	-0.304	-0.838	0.231
artistliving1:year	-0.052	-0.096	-0.008
diff_origin1:Surface	0.000	0.000	0.000
diff_origin1:portrait1	0.571	-0.294	1.436
diff_origin1:still_life1	-1.131	-1.803	-0.458
diff_origin1:prevcoll1	0.093	-0.763	0.948
endbuyerC:Surface	-0.001	-0.002	0.001
endbuyerD:Surface	-0.001	-0.002	0.001
endbuyerE:Surface	-0.001	-0.002	0.001
endbuyerU:Surface	-0.001	-0.002	0.001
endbuyerX:Surface	-0.001	-0.002	0.001
paired1:endbuyerC	-1.416	-2.846	0.013
paired1:endbuyerD	-1.035	-2.456	0.385
paired1:endbuyerE	-0.672	-2.161	0.816
paired1:endbuyerU	-0.894	-2.355	0.566
paired1:endbuyerX	-1.043	-2.495	0.410
endbuyerC:year	-0.006	-0.168	0.155
endbuyerD:year	-0.066	-0.227	0.095
endbuyerE:year	0.039	-0.124	0.203
endbuyerU:year	-0.042	-0.205	0.121
endbuyerX:year	-0.060	-0.220	0.101
portrait1:authorstyle1	0.036	-1.835	1.907
Interm1:lrgfont1	-0.273	-0.752	0.206
paired1:lrgfont1	-0.581	-1.064	-0.097
paired1:subjectArch	0.544	-0.633	1.721
paired1:subjectBattle	-0.741	-2.008	0.527
paired1:subjectBuste	1.455	0.215	2.694
paired1:subjectFruit$Flower	0.166	-0.944	1.275
paired1:subjectMarine	-0.717	-1.936	0.503
paired1:subjectother	-0.170	-0.986	0.646
paired1:subjectPaysage	0.087	-0.746	0.919
paired1:subjectPeople	-0.156	-1.031	0.718
paired1:subjectPortrait	0.093	-0.984	1.170
paired1:subjectSaint	0.343	-0.598	1.284
paired1:subjectSujet	-0.333	-1.564	0.897
paired1:discauth1	-0.350	-0.983	0.283
paired1:authorDavid Teniers	-0.964	-2.428	0.499
paired1:authorFrancois Boucher	0.932	-0.701	2.565
paired1:authorFrench	0.525	-1.707	2.757
paired1:authorGaspard Netscher	0.482	-1.693	2.657
paired1:authorGasparo Van Vitelle	-0.655	-2.608	1.297
paired1:authorNicolas Berghem	0.126	-1.761	2.014
paired1:authorNicolas Poussin	0.097	-2.648	2.842

	Coefficient	2.5%	97.5%
paired1:authorother	-0.115	-1.412	1.183
paired1:authorPhilippe Wouvermans	0.645	-0.927	2.217
paired1:authorRosalba Carriera	0.033	-1.800	1.866
finished1:discauth1	0.761	0.230	1.292
lrgfont1:discauth1	-0.768	-1.459	-0.078
dealerL:prevcoll1	-0.520	-1.746	0.706
dealerP:prevcoll1	-1.679	-3.280	-0.078
dealerR:prevcoll1	-0.590	-1.500	0.321
Interm1:prevcoll1	-0.488	-1.112	0.135

Variables:

a. The base variables we chose are: `Shape`, `school_pntg`, `dealer`, `Interm`, `paired`, `artistliving`, `diff_origin`, `endbuyer`, `finished`, `year`, `Surface`, `portrait`, `still_life`, `prevcoll`, `authorstyle`, `lrgfont`, `discauth`, `subject` and `author`.

b. The interactions we used 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`.

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:

a. Linear Model

The linear model from part I does a fairly good job in predicting. Therefore, after adding two more variables in the dataset, we decided to refine the linear model first. The plan was to add new features and interactions into the model, hoping to potentially explain more variation in the response variable. 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 predicting. However, the output from AIC contain too many interactions, which might lead to the problem of overfitting. Additionally, it contains some interactions with coefficients as NA, and some that do not make sense at all. Therefore I manually removed them and kept twisting around the rest features, which led to the best final linear model in terms of the lowest RMSE.

b. Tree Model

Since we have many categorical variables, as in nature, a tree-based model would be appropriate in addressing the interactions to explain the response variable. We tested two tree models: random forest and boosting (bagging does not work in our 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 do better job than the linear model from part I, but not as good as the linear model generated above.

c. Poisson&Negative Binomial Regression

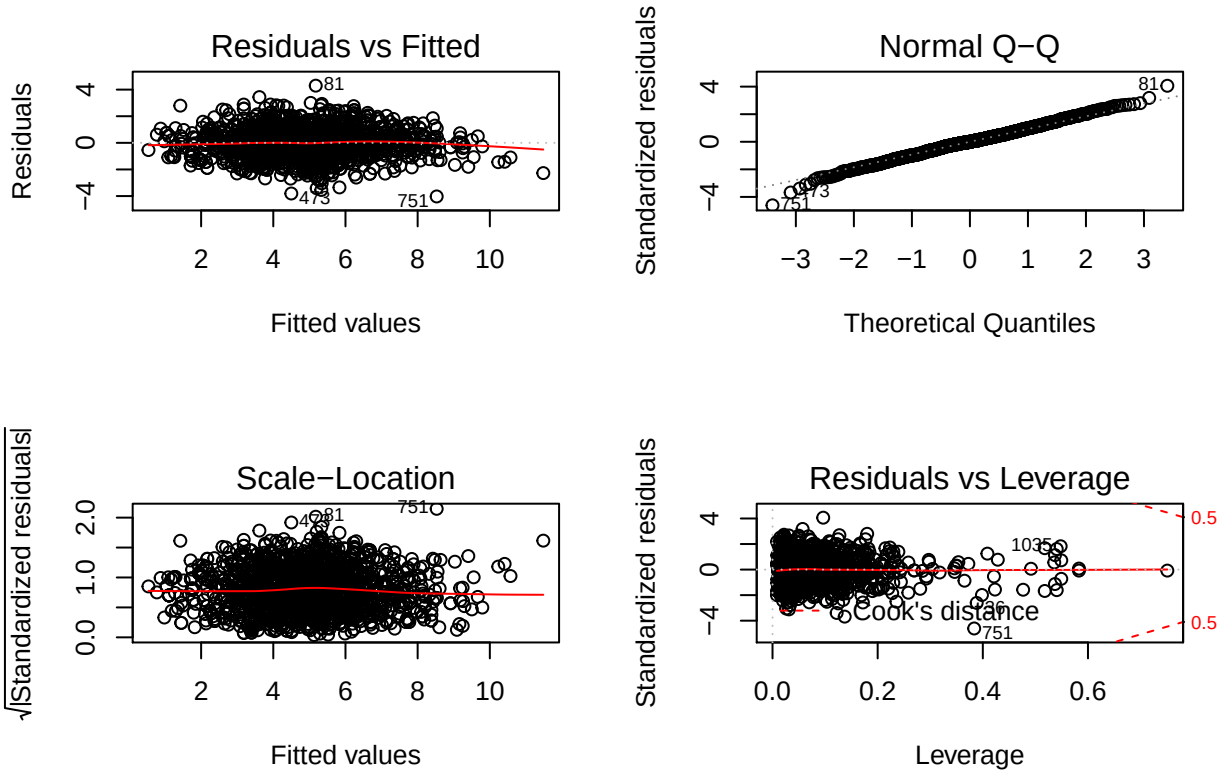
The nature of auction price is integer. Even though not strictly count data, but could be treated in such way so that we can use poisson&negative binomial regression. We trained the tested the poisson model first. With all the vairiables included (even with all the interactionss), the residual deviance is 100 times higher than the residual degrees of freedom. We concluded that the poisson model s not appropriate and proceeded with negative binomial model. The RMSE is not better than the linear model is part I (even with all the interactions) and there is still the problem of over-dispersion.

d. Xgboost ## Added some description here

Comparing all the models we fitted above, we conclude that the linear model has the best performance in terms of predicting (lowest rmse). The linear model was relatively more complex than the one in part I, resuting in some loss in predictability. However, it is still more interpretable than tree-based models. Therefore, we concluded that the best model is the linear model.

Residual: must include a residual plot and a discussion

The residual plot looks fairly good. There seems to have a pattern that the variance is slightly higher with fitted value around 5 and a little lower at the two tails. But in general, the constant vairance assumption is satisfied. The normality assumption is well satisfied from normal qq-plot, with several observations slightly scattered away on the two sides.



discussion of how prediction intervals obtained

Table 4: Prediction Interval

fit	lwr	upr
100.911	11.113	916.327
281.695	20.514	3868.191
188.818	19.938	1788.136
4661.882	404.135	53776.928
449.689	48.002	4212.741
298.025	28.963	3066.648
133.289	14.698	1208.732
3.297	0.334	32.561
107.448	11.447	1008.519
114.416	11.665	1122.194
29.444	3.097	279.882
459.767	50.263	4205.557
467.064	49.678	4391.264
1043.635	113.681	9580.940
16.832	1.839	154.053
87.815	9.714	793.853
2238.476	244.502	20493.821
1081.348	117.755	9930.014
104.930	9.576	1149.804
15.656	1.577	155.402

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, we just showed the prediction interval for the first 20 observations.

5. Assessment of the final model

Model evaluation: must include an evaluation discussion

The final model that is specified has generally good 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.

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 high leverage.

- magnitude of coefficients and standard errors

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.

Model testing : must include a discussion

- metrics from test set
- potential improvelemnts - k-fold cross validation due to uncertainty in makeup of test set. different test/validation sets may yield different model results

After selecting the final model and predicting with the test dataset, we looked at a couple of metrics to evaluate the performance of the model. The two most important evaluations of the model are its rmse value, which represents the goodness of fit of the model, and its bias, which measures the tendency for the model to systematically over- or underestimate.

In terms of the final model that we fit, we have the rmse value at 1210.42 and bias at 184.08. Since the method that we end up using is a linear model, whose ultimately purpose and model selection criteria is not to miniimize the rmse value, the rmse value that we end up with, although is not the best, is reasonaly good. Also, since we used stepwise AIC as the model selection criteria, it gives us a model that has better performance at predicting (comparing to BIC who leads to a model closer to the "true model"), the resulting model would have a higher rmse value than if we were to choose a different cirteria.

Lastly, since we aim at balancing the simplicity, interpretability and performance of the model, as well as the tidyness of the , certain levels of categorical variables are merged and interactions were removed from the

model, even though the stepwise function suggested to include them. This could also negatively affect the rmse and bias of the final selected model, but would also add up the difficulty of using and understanding the model, thus we chose to slightly sacrifise the precision.

Model result: must include a selection and discussion of the top 10 valued paintings in the validation data.

price	rownum
55503.3	2065
34130.0	2543
30240.9	2066
27905.0	2538
24234.8	1071
18021.1	2584
17564.4	2517
15974.5	2477
13626.9	2022
10163.9	2528

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 of these shared features and the level/status these 10 paintings are at, after taking into account the related interaction terms, lead to a positive correlation with the price. Therefore, the model ended up predicting high prices for these paintings. It is interesting that, the most expensive five paintings predicted all come from the same author Nicolaes Berchem Pieterszoon. This phenomenon further confirms the consistency of the final model in terms of predicting, because usually paintings from those artists who tend to have highly valued pieces of arts in auctions would all be pricy, and it is common for the same dealer and endbuyer to purchase paintings from the same artist, and the features of the artists' paintings are similar.

6. Conclusion (10 points): must include a summary of results and a discussion of things learned. Optional what would you do if you had more time.