Part-II Complex Model

Chenxi Wu, George Lindner, Qianyin Lu, Yi Mi

12/14/2019

Introduction

Our team of esteemed statisticians was recently hired by a prestigious art historian for a consulting project. We were asked to help build a predictive model in exchange for an A on our STA 521 Final Exam. After much discussion, our team accepted the historian's offer. We were given the task of predicting paintings' selling prices at auctions in 18th century Paris. To accomplish this, we used a dataset containing information about each painting's buyer, seller, painter, and characteristics of the painting.

There were two primary objectives in our analysis:

1) To determine which variables (or interactions) drove the price of a painting.

2) To determine which paintings were overpriced or and which were underpriced.

The first objective could be accomplished through EDA and modeling. Getting to know the dataset through EDA helps our team identify relationships in the data and develop a sense of which variables might be important for prediction. This developed intuition of the data helps our team begin modeling the logprice variable. After an extensive modeling process, we can report with confidence which variables are drivers of a painting's selling price.
When we fit the final model, we can calculate how far each painting's selling price deviates from our prediction. Positive residuals indicate that a painting sold for more than we think it is worth. The opposite goes for negative residuals. Therefore, we can achieve our second goal through a residual plot analysis of our model.
We had 1,500 observations to train the model on, along with 750 observations held out as a testing set. There was a total of 59 variables in the dataset, both categorical and continuous.

2. Exploratory Data Analysis:

Initial Data Cleaning

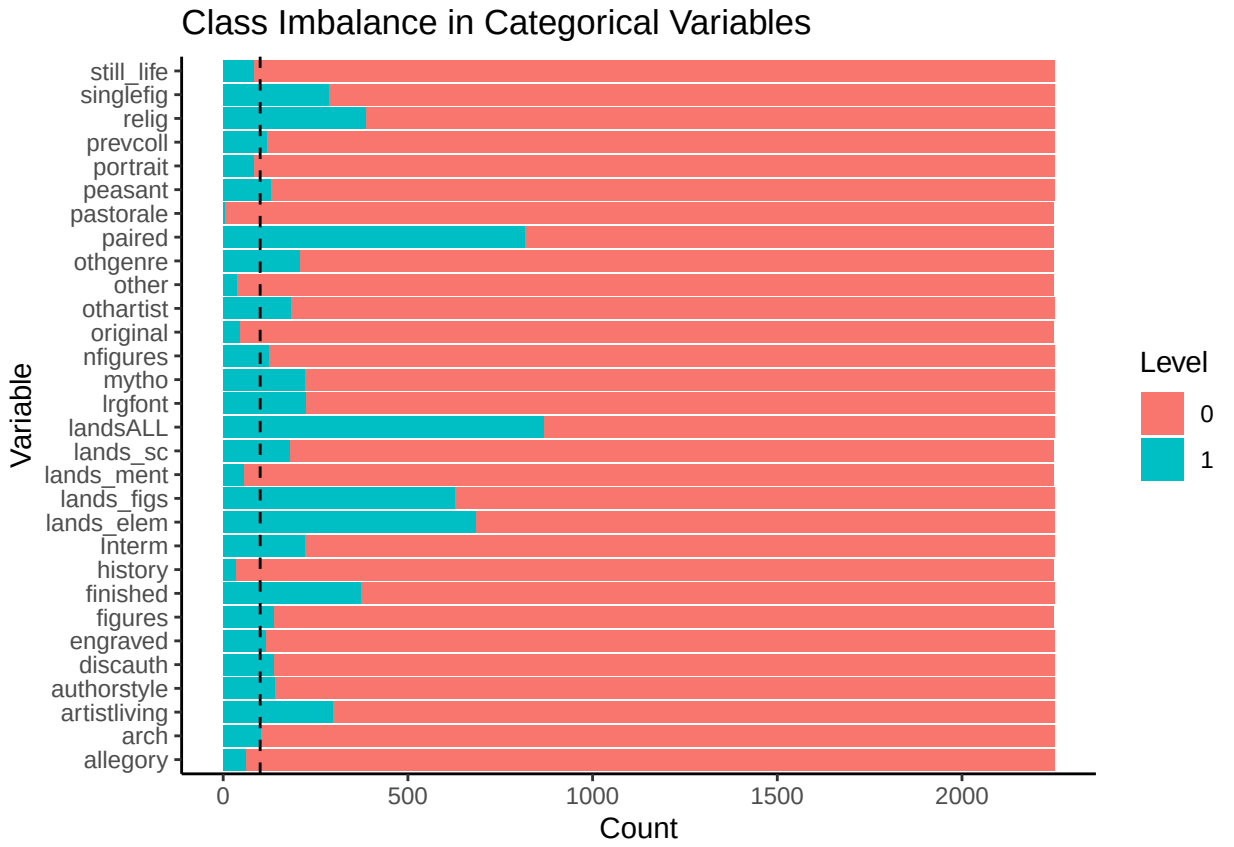
We began our data cleaning process by reading the codebook for a better understanding of what each variable in the data represented. Several predictors in the dataset were redundant and therefore removed to avoid high correlation among the predictors. Examples of this include the variable *sale*, which is a combination of *dealer* and *year*. Additionally, there were other predictors that we deemed would not be useful for prediction, such as *count* which was 1 for every observation, or *subject* which was a short description of the content in the painting. We simplified the data by eliminating unnecessary predictors. We also noticed that there are variables that record similar information, such as *figures*, *nfigures* and *singlefig*, for simplicity, we treated *nfigures* as binary variables and plotted boxplots of the three variables against the response (See Appendix). We decided to only include *nfigures* in our model building.

We then check on the emprirical distribution of the response variable. There are 2 variables, *logprice* and *price*. From the histogram (See Appendix), we can see that *logprice*, which is the logarithm of *price*, is more normally-distributed. Consider the normality assumption of linear regression, we will use *logprice* as the response variable.

Categorical Variables

We recoded each categorical variable to be a factor. We created a visualization of the binary categorical variables to observe the balance between classes below.

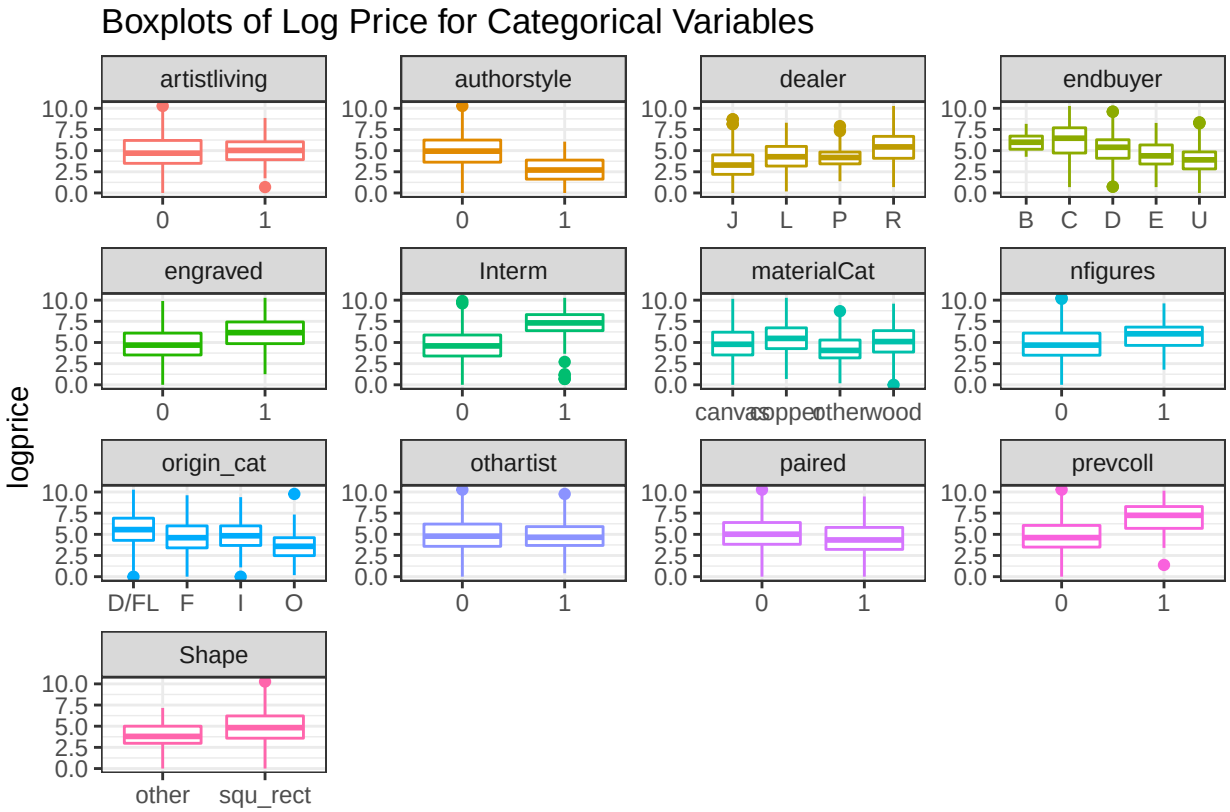
Plot 1

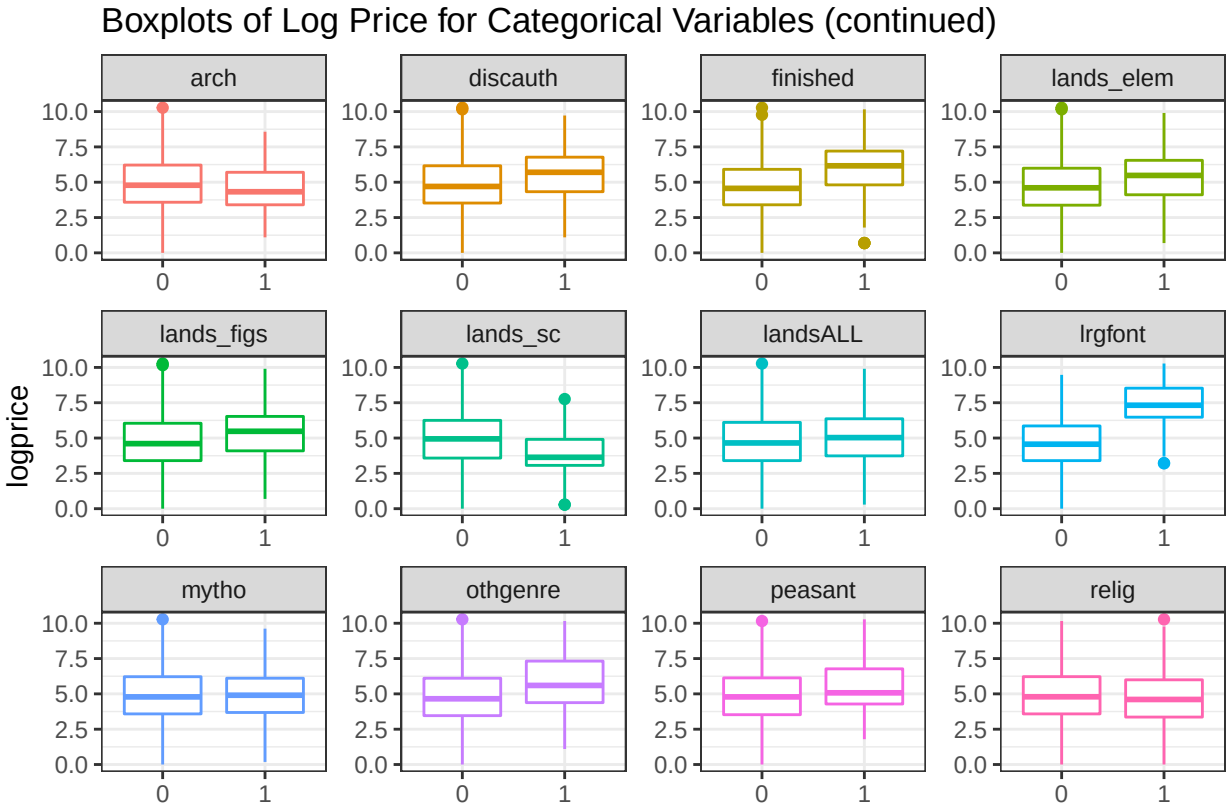


Imbalanced classes can lead to poor β estimates if the underrepresented class does not have enough data. This was our motivation to remove any variable that had less than an arbitrary 100 observations in a class, which is denoted by the dotted black line in our visualization above.

To identify important categorical variables, we created a boxplot for each variable that compared the distribution of *logprice* over every level of the factor. The results are shown below.

Plot 2


Boxplots of Log Price for Categorical Variables
logprice
artistliving
authorstyle
dealer
endbuyer
engraved
Interm
materialCat
nfigures
origin_cat
othartist
paired
prevcoll
Shape
0
1
J
L
P
R
B
C
D
E
U
canvas
copper
otherwood
D/FL
F
I
O
other
squ_rect


Boxplots of Log Price for Categorical Variables (continued)
logprice
arch
discauth
finished
lands_elem
lands_figs
lands_sc
landsALL
lrgfont
mytho
othgenre
peasant
relig
0
1

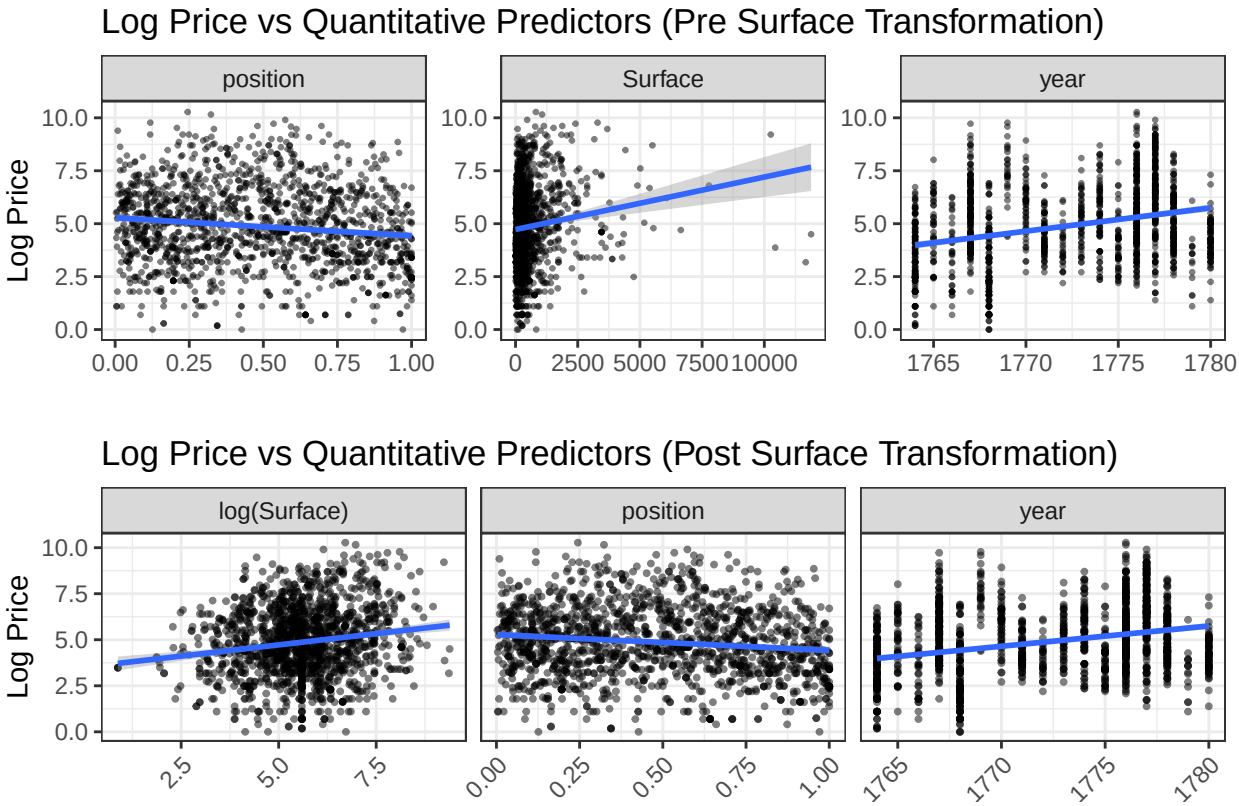
The boxplots above help us identify which variables could be important in predicting a painting's price. They also help us in our variable selection process by displaying variables that have similar prices in all of their categories. After inspecting the boxplots, we determined that *mytho*, *landsALL*, *relig*, and *othartist* were not useful for prediction. Variables that may be important include, but are not limited to, *lrgfont*, *Interm*, *authorstyle*, and *prevcoll*.

Quantitative Variables

There are also quantitative variables in our data that could be used for prediction. Like the categorical variables, many of these predictors were redundant. For example, we were given the surface area of a painting. Additionally, we were given a variable for surface area if the painting was round and a surface area variable if the painting was rectangular. We also were given the height, the width, and the diameter of the painting. We determined that all this information could be condensed to a single variable, *Surface*.

There were missing data in *Surface* that we had to address. Surface area intuitively seems like it could drive the price of a painting, so we had to develop a strategy for handling the missing observations. With the help of the plot below, we determined that imputing the median surface area size of the dataset would be a good estimation for missing values. Since the distribution of *Surface* is skewed, we wanted an imputation strategy that would be robust to outliers. Thus, we opted for the median over the mean.

Plot 3



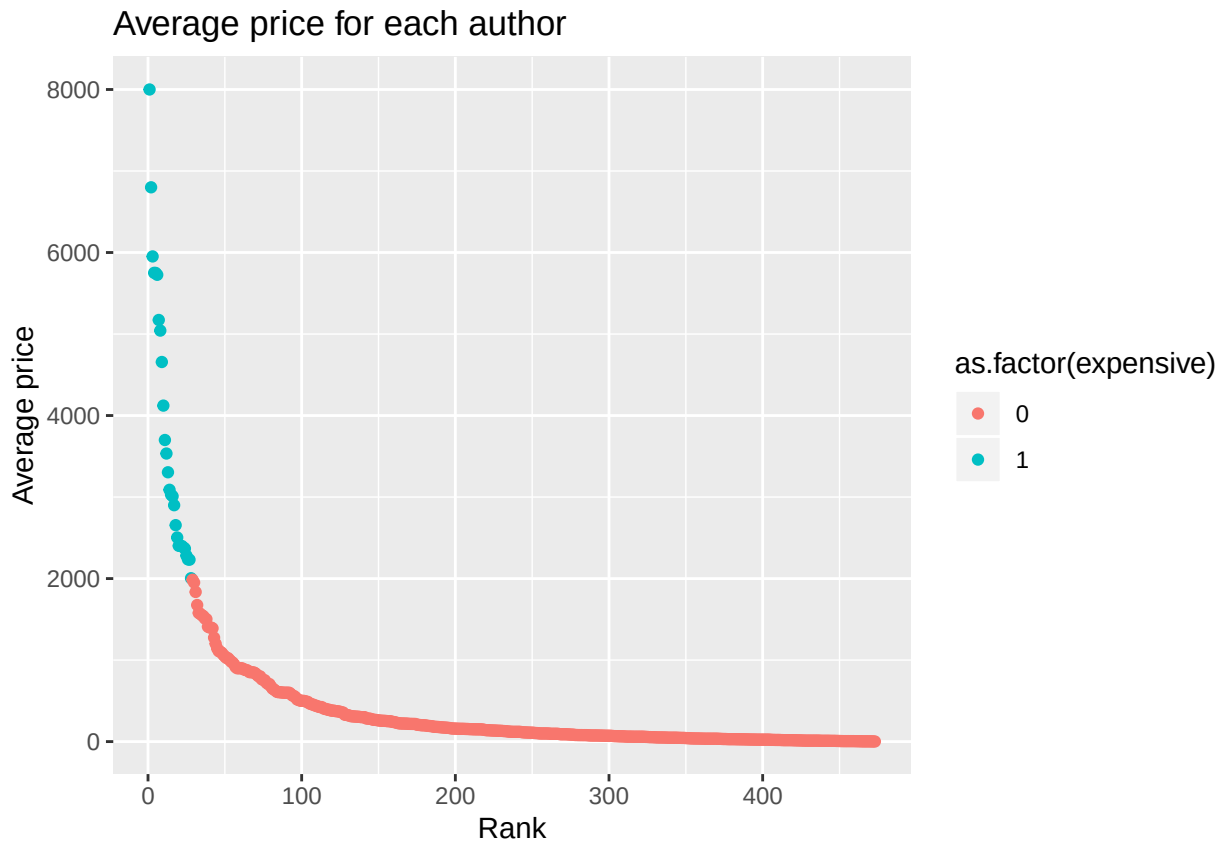
We created scatterplots to observe the relationship between our three quantitative predictor variables and the log price of a painting. The distribution of *Surface* was skewed right and a log transformation was necessary. We plot the relationship of logprice and the log transformed Surface column in the lower graph.

Additional EDA after Part 1

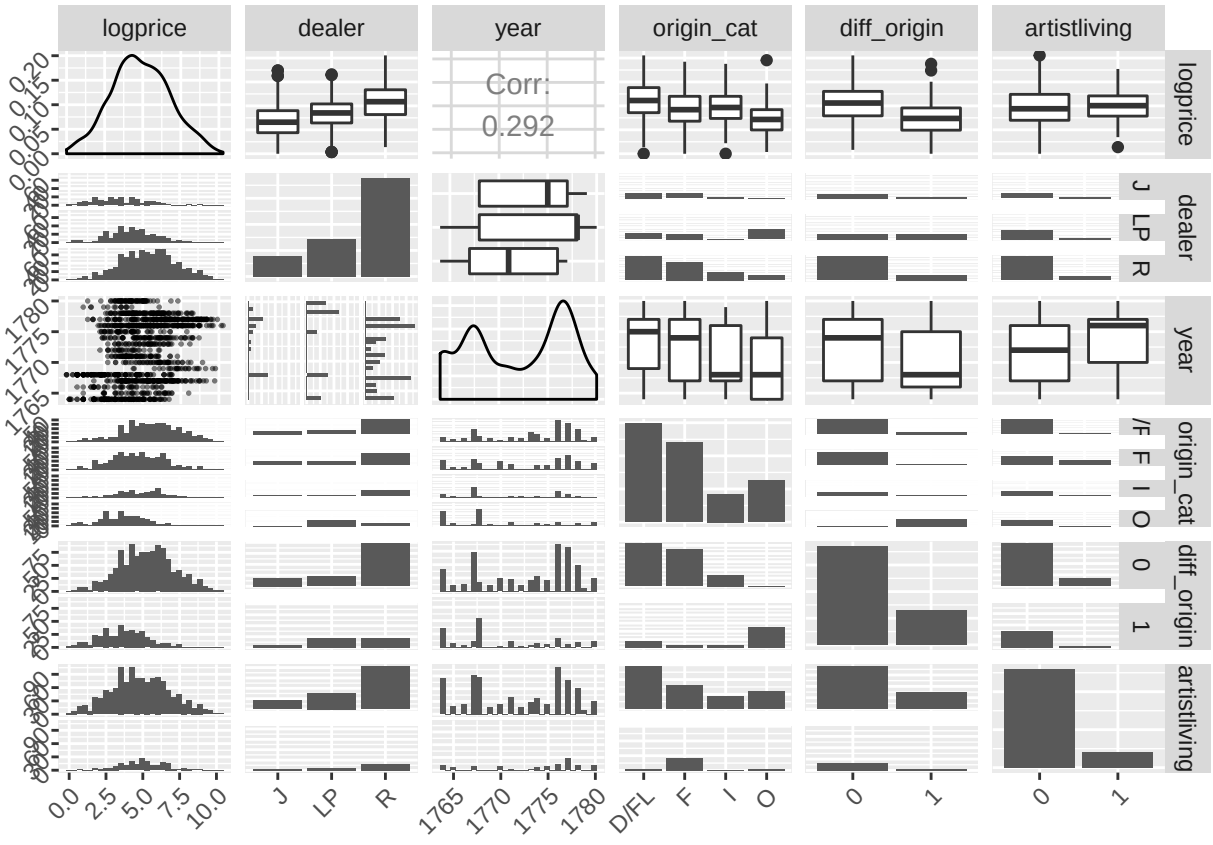
Based on EDA of Part-I, included above, we improved our data manipulation on our data as follows:

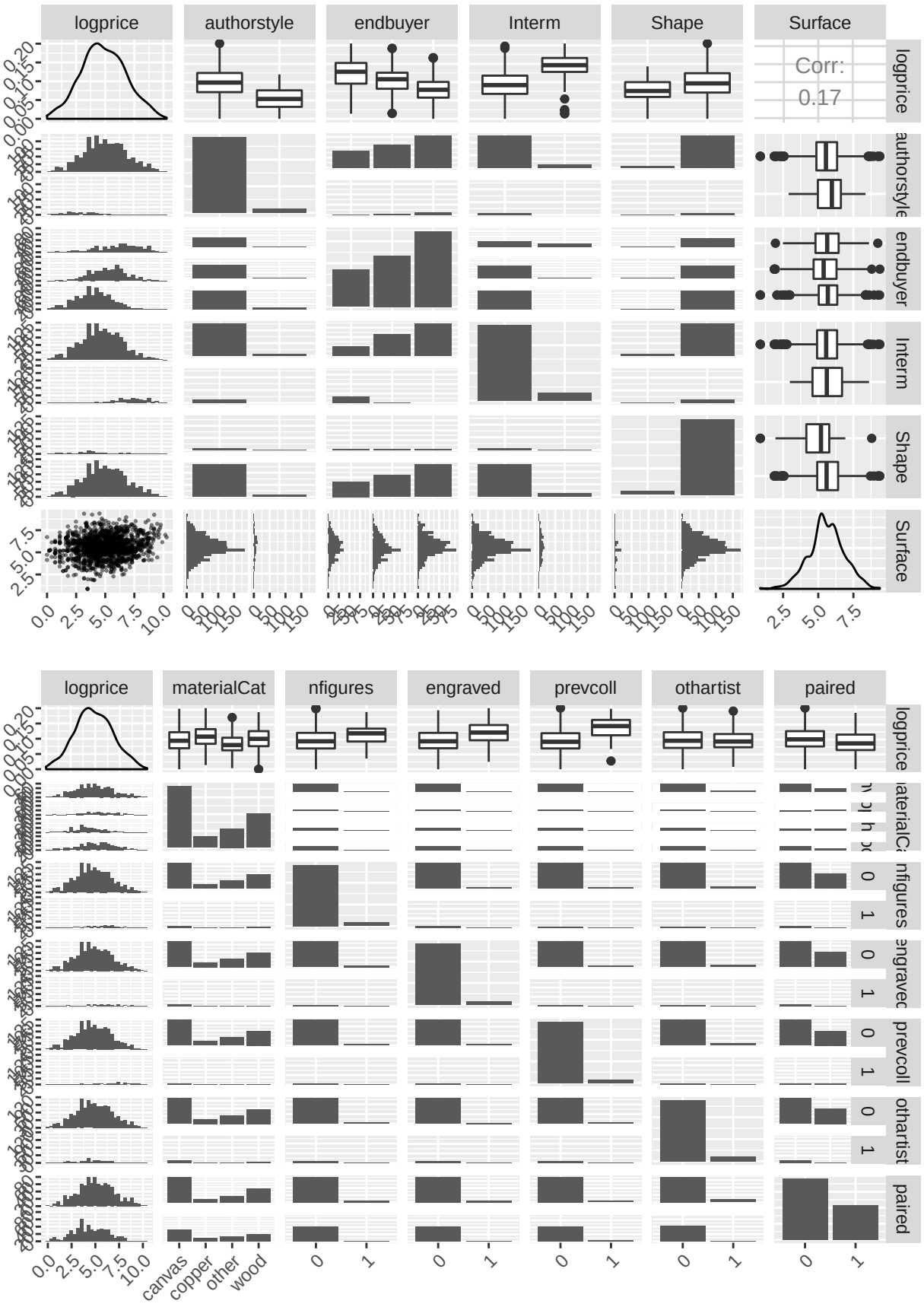
- *position* has values greater than 1 which should be data entry errors, we divided them by 100 to get the right value.
- The original dataset contains lots of missing values and NA's, like *winningbiddertype*, *endbuyer*, *authorstyle*, *Interm* and *type_intermed*, we filled the missing values with "U", "Unknown" or 0 according to the description of codebook.
- Most of observations for *Shape* are "squ_rect", so we regroup other shapes to "other". After testing the average *logprice* of "other" and the missing ones, we decided to recode the missing values to "other" since they have similar average *logprice*. For same reasoning we recoded the missing values in MaterialCat to "other".
- To alleviate the class imbalance problem of *school-pntg*, *origin_cat*, *mat* and *material*, we regrouped levels with fewer observations to larger levels.
- We transformed *nfigures* into a binary variable where values other than 0 are set to 1 since the empirical distribution of *nfigures* is extremely skwewed and most of the values gather around 0.
- In Part-I we imputed the NA's in *Surface* to median value of *Surface*. Here we tried more advanced methods by regressing other variables on *Surface* to see the correlations. We found out that *Surface* was correlated with *MaterialCat* and *relig*, from which we devided the data into 8 groups and imputed median value for each group respectively. We tested the efficiency of the new imputation and the result showed that *Surface* has more explanation power than before.
- In Part-I we discarded the variable *authorstandard* which can be a strong predictor. Here we cleaned *authorstandard* so it contains fewer unique values. We computed the average price for each author and ploted them in a descending order (See plot below). The plot showed that the relationship between author and price is significant. So we created a binary variable *expensive*, we set the authors with high average price to 1 and the others to 0. The variable we built actually captures a significant amount of variation in the response variable. the regression of *expensive* on *logprice* achieved an $R^2 = 0.157$.
- To avoid overfitting, we regrouped *dealer* and *endbuyer* into three levels respectively. Specifically, we combined 'L' and 'P' in *dealer* and 'E' and 'U' in *endbuyer*.

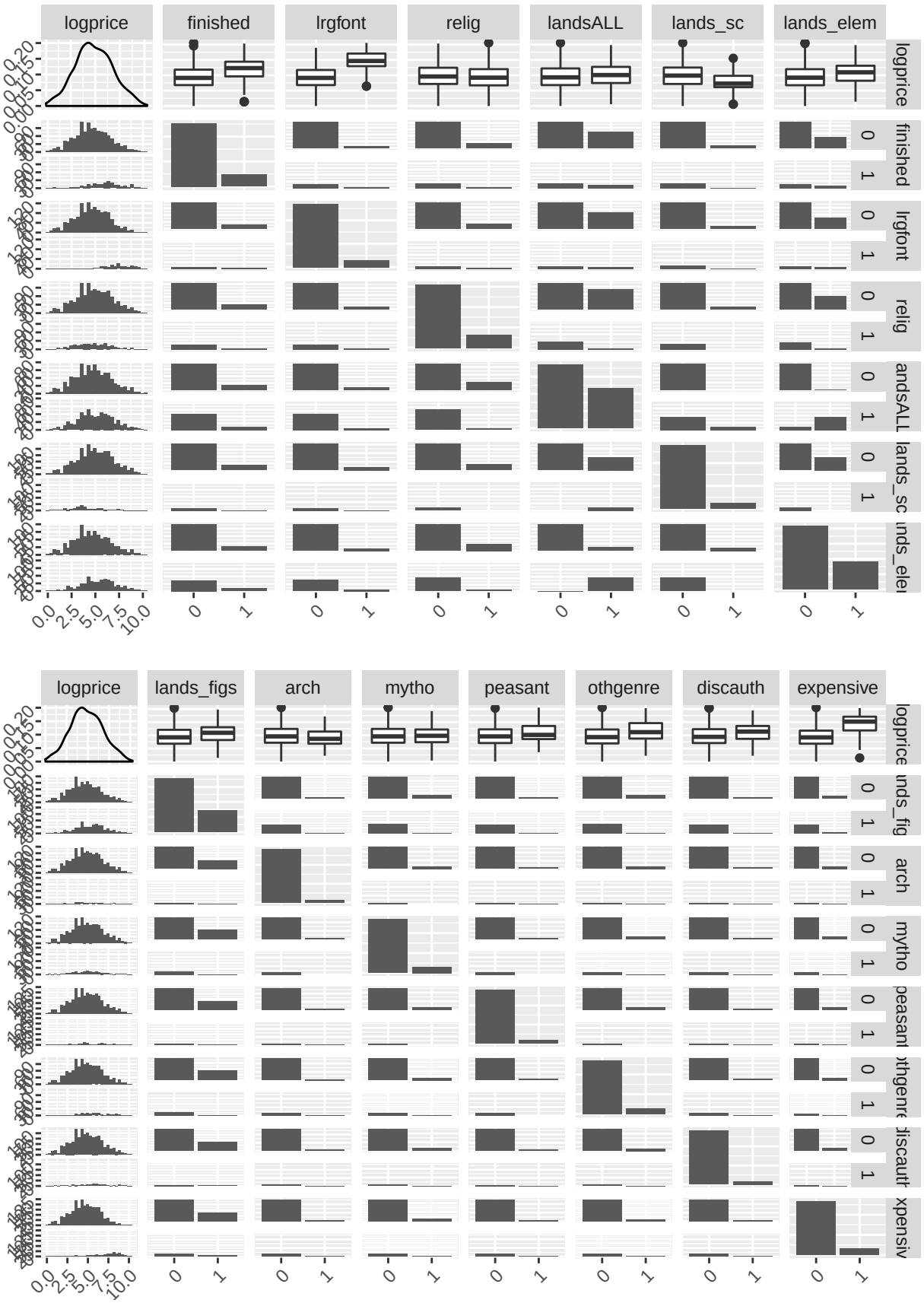
Plot 4: Splitting Authors Based on Price



Plot 5: Pairwise Plots




logprice
authorstyle
endbuyer
Interm
Shape
Surface
Corr:
0.17
materialCat
nfigures
engraved
prevcoll
othartist
paired
canvas
copper
other
wood


logprice
finished
lrgfont
relig
landsALL
lands_sc
lands_elem
0.0
2.5
5.0
7.5
10.0
0
1

logprice
lands_figs
arch
mytho
peasant
othgenre
discauth
expensive
0.0
2.5
5.0
7.5
10.0
0
1

For EDA in Part-II, we added pairwise plot to have a general view of the interactions of all the variables. Here we listed some interesting findings.

- The continuous variable *year* and *Surface* seem to be positively correlated with *logprice*. Additionally, there seems to be a non-linear relationship between *year* and *logprice*.
- The pairwise plot of *year* and other categorical variables revealed that there might be interaction effect between these variables, which we should consider in model building.
- The interactions between categorical variables is not that significant due to **class imbalance**. There are simply not enough observation for most of the categorical variable interaction.
- Based on the plot, we could conclude that the most important predictors are: *year*, *dealer*, *origin_cat*, *diff_origin*, *expensive*, *authorstyle*,*endbuyer*, *Interm*, *Surface*, *materialCat*, *nfigures*, *engraved*, *prevcoll*, *paired*, *finished*, *lrgfont*, *lands_sc*, *lands_elem*, *othgenre*, *discauth*.

Discussion of Preliminary Model

After the test data was updated at 11 P.M. on December 12th, we went back to our preliminary model to check our true results. It turns out that this linear regression model was actually achieving 95.6% coverage instead of the mentioned 65% coverage in our Part I write-up. The bias was also significantly lower than we thought, coming in at 120.3. Our RMSE was still large, though, resulting in a score of 2360.

Table 1: Results from Preliminary Model

Bias	Coverage	MaxDeviation	MeanAbsDeviation	RMSE
120.29	95.6	52516.79	551.74	2363.27

Note:
Summarized from Werker

This model has low bias and high variance, meaning that we overfit the data. Coverage is sufficient so we want to focus our attention on improving the RMSE. This can be achieved through the **bias-variance trade-off**. We can significantly reduce the variance if we induce a little more bias into our model, thus improving our RMSE score.

The mean deviation was 551.74 but the max deviation was over 50,000. Our model is doing a good job on most predictions, but there are a few predictions that are extremely off, inflating the RMSE score. Our goal moving forward is to improve on these extreme cases and to introduce a little more bias into the model to produce a lower RMSE.

Development of the final model

We tried several complex models to better depict the behaviour of the response variable. The findings of those models are summarised as below.

Random Forest

Since we have many variables and the interactions among them can be involved, a tree model seems to be appropriate for the setting. To alleviate the unstability of single tree models, we used random forest method to achieve more robust estimation. We select *year*, *dealer*, *origin_cat*, *diff_origin*, *expensive*, *authorstyle*,*endbuyer*, *Interm*, *Surface*, *materialCat*, *nfigures*, *engraved*, *prevcoll*, *paired*, *finished*, *lrgfont*, *lands_sc*, *lands_elem*, *othgenre*, *discauth* as predictors based on the EDA above. The 10 most important variables are *experience*, *year*, *Surface*, *endbuyer*, *dealer*, *materialCat*, *origin_cat*, *paired*, *Irgfont* and *finished*. Below is the important variable plot and the 5 least important variable table. We will discarded these 5 least important variables in further modeling.

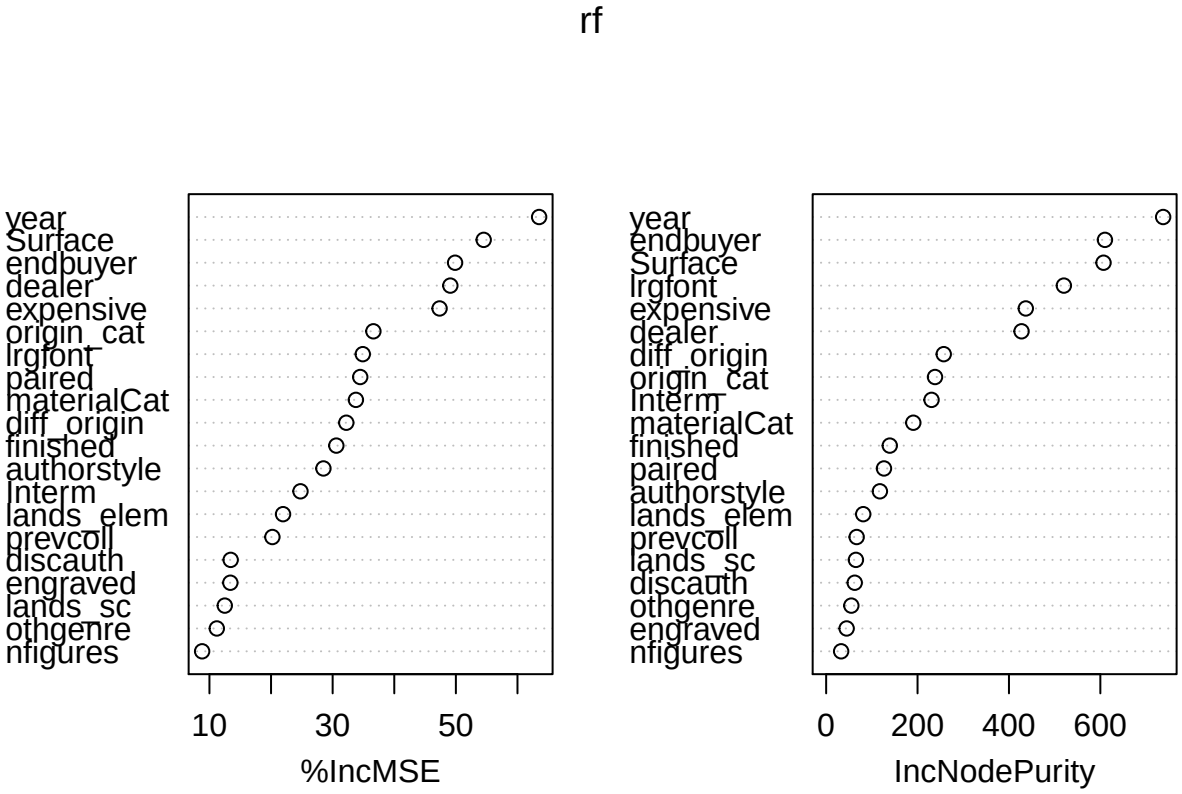


Table 2: Least 5 important variables of RF

	Overall	Vars
13	20.217	prevcoll
20	13.444	discauth
12	13.395	engraved
17	12.488	lands_sc
19	11.196	othgenre

To assess the performance of the random forest model, we first evaluated it using training set which achieved a training RMSE of 1063.14634. However when we used it for test set, the prediction contains only a point estimate instead of a prediction interval. We tried to compute the interval using the quantile method, but the coverage is not ideal, which might due to narrower interval. So we move on to other variable selection method like Bayesian Model Averaging.

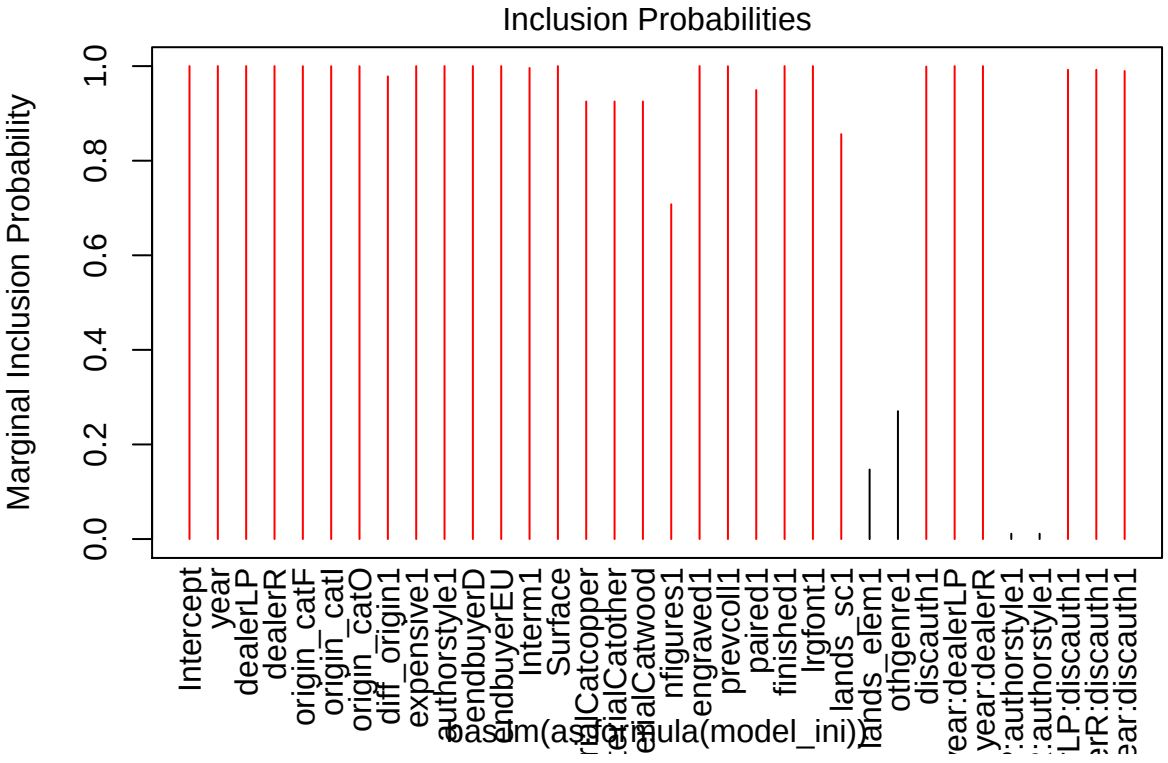
Lasso

Apart from including random forests in our model development, we also put consideration into Lasso since this model is suitable for both variable selection and preventing overfitting through shrinkage. However, due to the fact that we have most of variables as categorical, some with multiple levels, it is hard to decide whether to normalize these predictors before modeling because if we do so, the result will depend on class prevalence and for multilevel predictors, we need to regroup them. The result will vary depending on our reference level and generate more difficulty for interpretation as well. In addition, similar to trees, Lasso doesn't have an existing prediction interval and we will need to use bootstrap to obtain such interval. Thus, to make our model easier to understand and more convenient for predictions, we decide to move on and do not include Lasso in the model-building process.

Bayesian Model Averaging

From the analysis so far our main problem is **overfitting**. This might be improved with Bayesian Model Averaging (BMA) which is an application of Bayesian inference to the problems of model selection, combined estimation and prediction that produces a straightforward model choice criteria and less risky predictions. [1]

In addition to the variables we used in the random forrest model, we also added interactions based on the p-value of these interactions (See Appendix for the summary of the full model), namely *dealer* with *year*, *authorstyle* and *discauth*, *year* and *discauth*. The results can be summarised as follows:

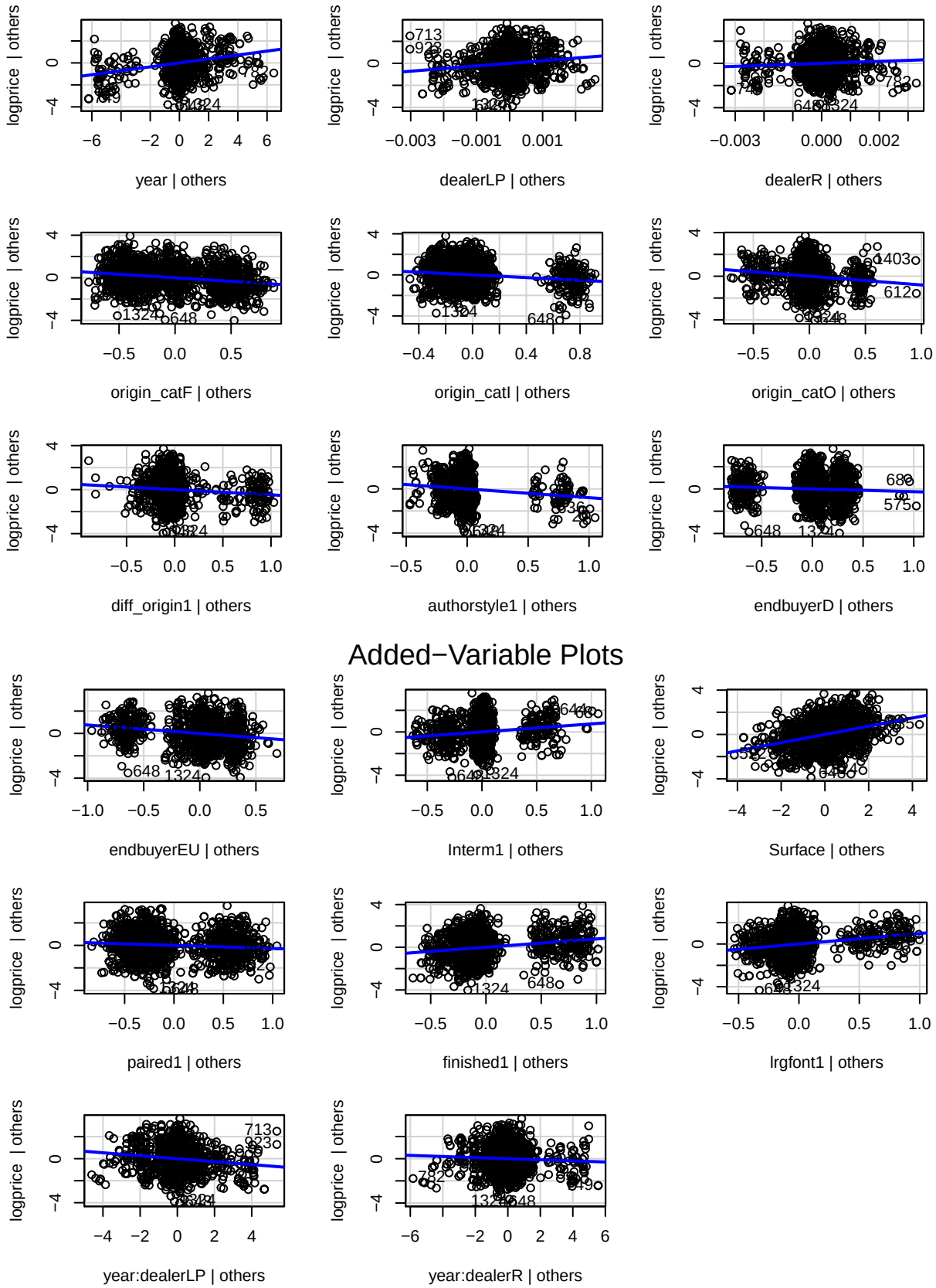


From the marginal inclusion probability plot, we should exclude *materialCat*, *nfigures*, *lands_elem*, *authorstyle:dealer* since their margnial inclusion probability is less than 0.5. ### Linear Regressin and Further Variable Selection

Table 3: Coefficient Summary of Best BMA model

variable	coef	lwr	upr
discauth1	348.833	180.639	514.784
dealerLP	217.078	131.063	303.933
dealerR	74.513	-4.162	154.854
Intercept	4.868	4.812	4.924
expensive1	1.080	0.876	1.276
engraved1	0.759	0.478	1.014
lrgfont1	0.682	0.458	0.905
prevcoll1	0.654	0.385	0.926
finished1	0.582	0.400	0.756
Interm1	0.528	0.278	0.782
materialCatcopper	0.365	0.000	0.594
Surface	0.346	0.285	0.407
nfigures1	0.273	0.000	0.582
year	0.170	0.129	0.213
materialCatwood	0.162	0.000	0.325
othgenre1	0.066	0.000	0.348
lands_elem1	0.020	0.000	0.178
dealerR:authorstyle1	-0.008	0.000	0.000
dealerLP:authorstyle1	-0.009	0.000	0.000
year:dealerR	-0.041	-0.087	0.002
year:dealerLP	-0.122	-0.169	-0.072
year:discauth1	-0.196	-0.292	-0.106
materialCatother	-0.220	-0.392	0.000
paired1	-0.220	-0.338	0.000
endbuyerD	-0.228	-0.413	-0.036
lands_sc1	-0.313	-0.544	0.000
diff_origin1	-0.413	-0.660	-0.177
origin_catF	-0.424	-0.584	-0.261
origin_catI	-0.447	-0.657	-0.237
origin_catO	-0.607	-0.906	-0.298
endbuyerEU	-0.726	-0.912	-0.531
dealerLP:discauth1	-0.796	-1.983	0.437
authorstyle1	-0.871	-1.171	-0.588
dealerR:discauth1	-2.410	-3.350	-1.544

To be thorough with our analysis, we decided to fit one more linear regression model using what we learned from the Random Forest and BMA models. We cross referenced the important variables that the RF and BMA models agreed on, and used these in our linear model. However, our overfitting problem still exists so we agreed upon observing Added Variable plots for further variable selection. These plots show us the relationship between the response variable and one of the predictors in the regression model, after controlling for the presence of other predictors.

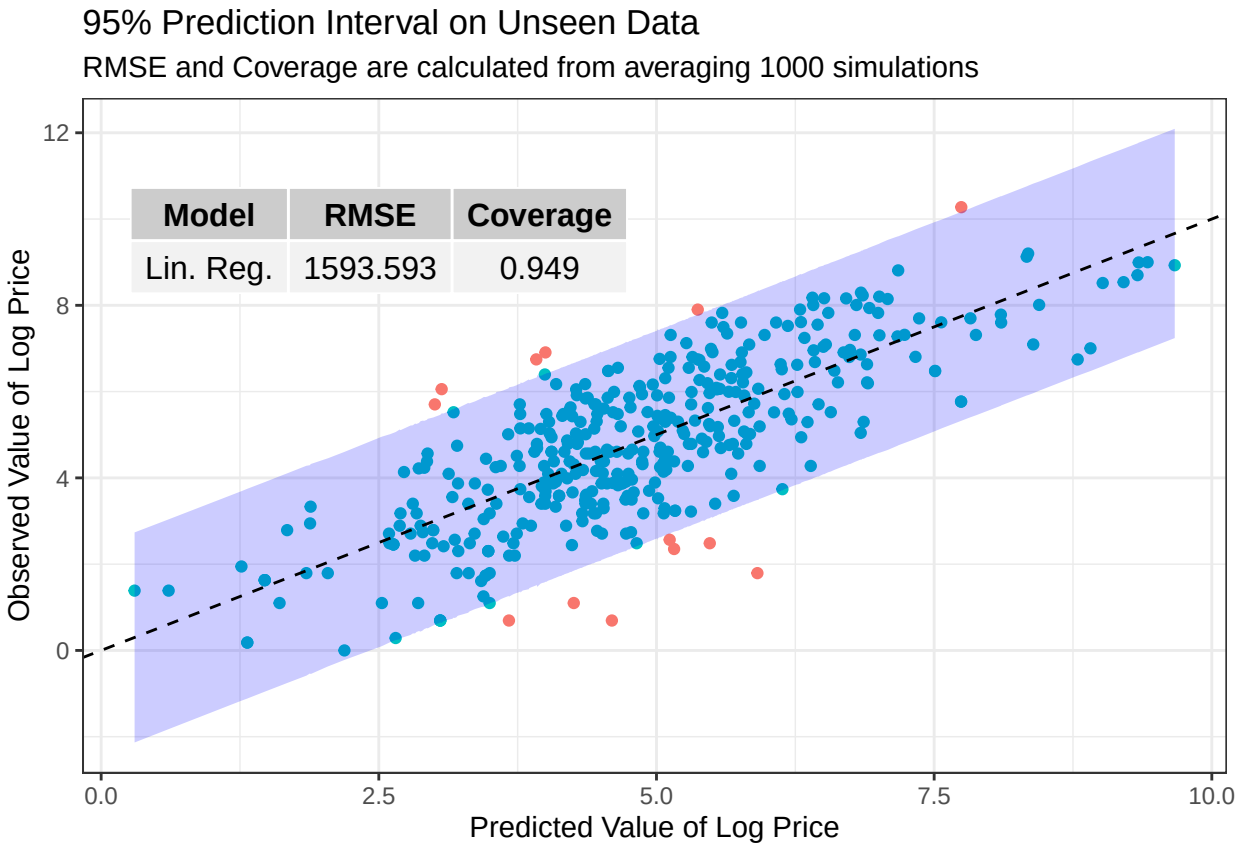


From the Added variable plots above, we observed that the regression line is nearly flat with variables *paired*,

origin_cat and *year*:*dealer*, so we deleted these variables to refit the linear model.

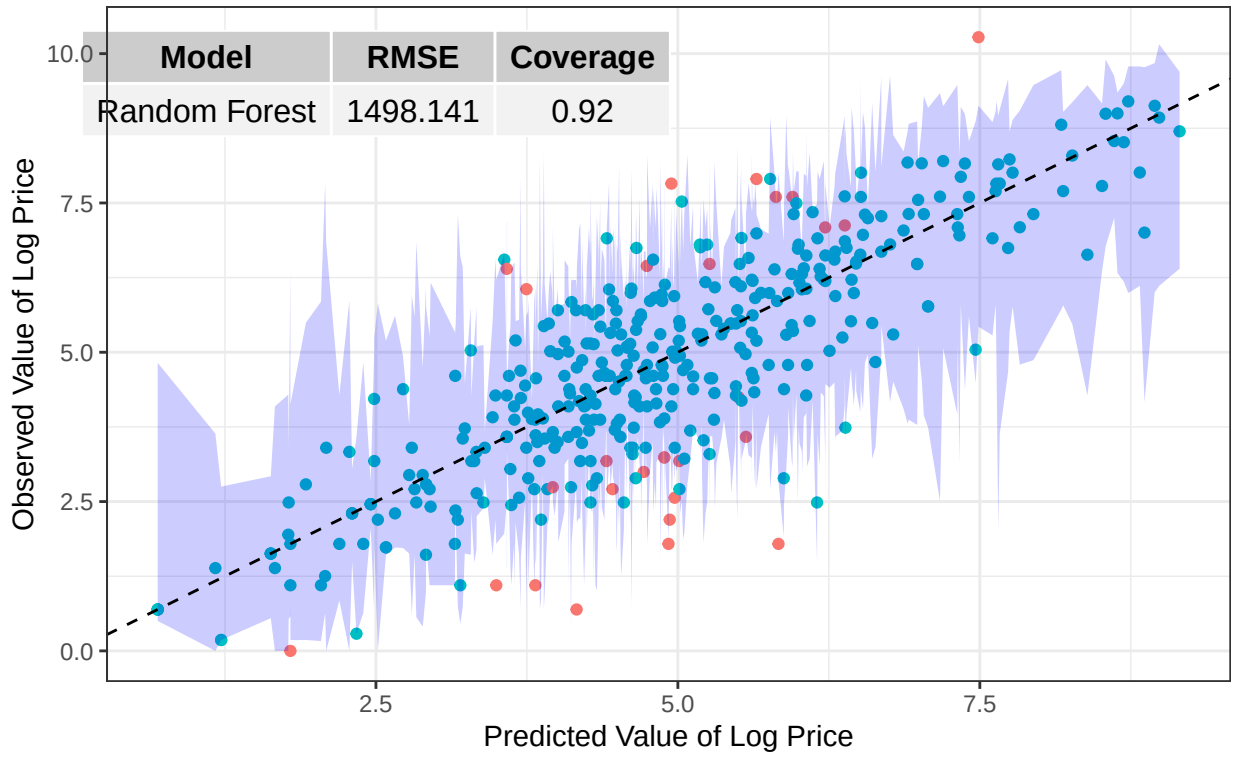
Model Selection

Now that we have 3 models to compare, we created a series of functions that would sample data from the training set, train the model on this data, and calculate RMSE and Coverage on the remaining unseen validation observations. Since the linear model is quick to fit, we ran this sampling simulation 1000 times to obtain more stable results on the out-of-sample validation data. We also ran 50 simulations on the Random Forest algorithm and 5 BMA simulations. The results from the models on unseen validation data help us choose our final model.



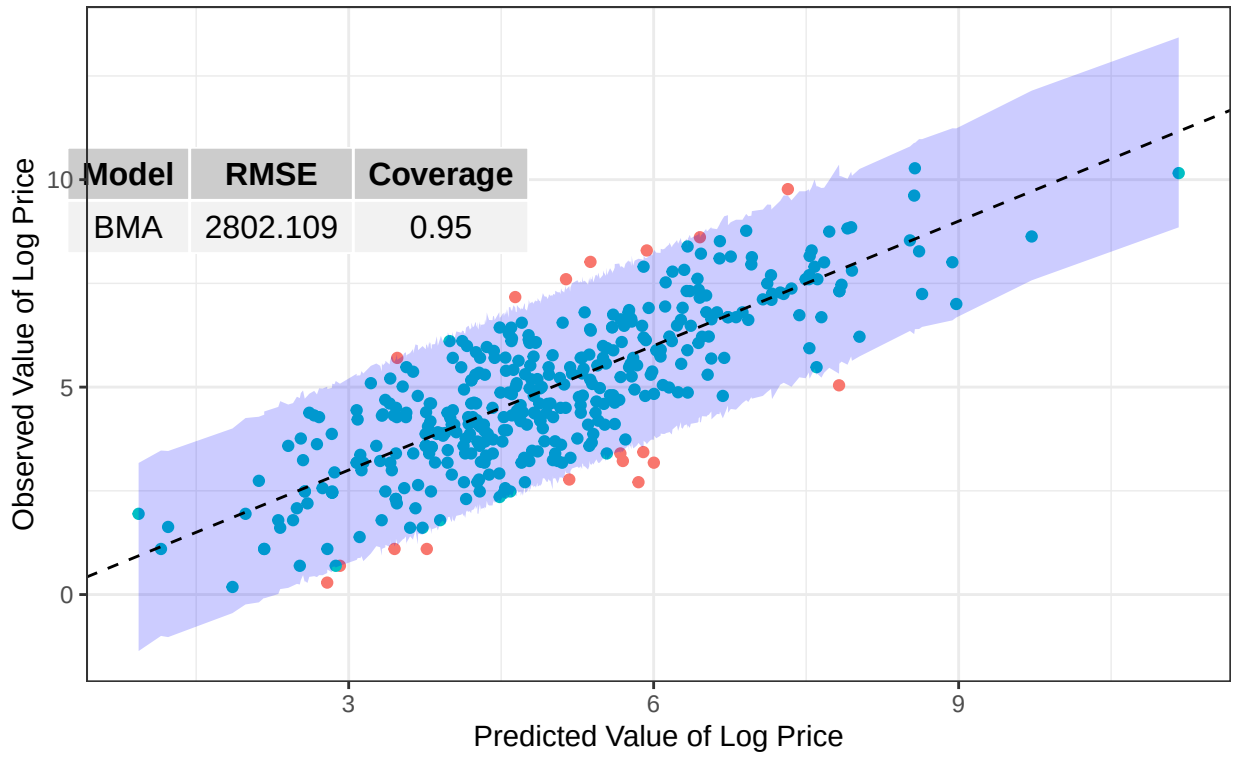
95% Prediction Interval on Unseen Data

RMSE and Coverage are calculated from averaging 25 simulations



95% Prediction Interval on Unseen Data

RMSE and Coverage are calculated from averaging 5 simulations

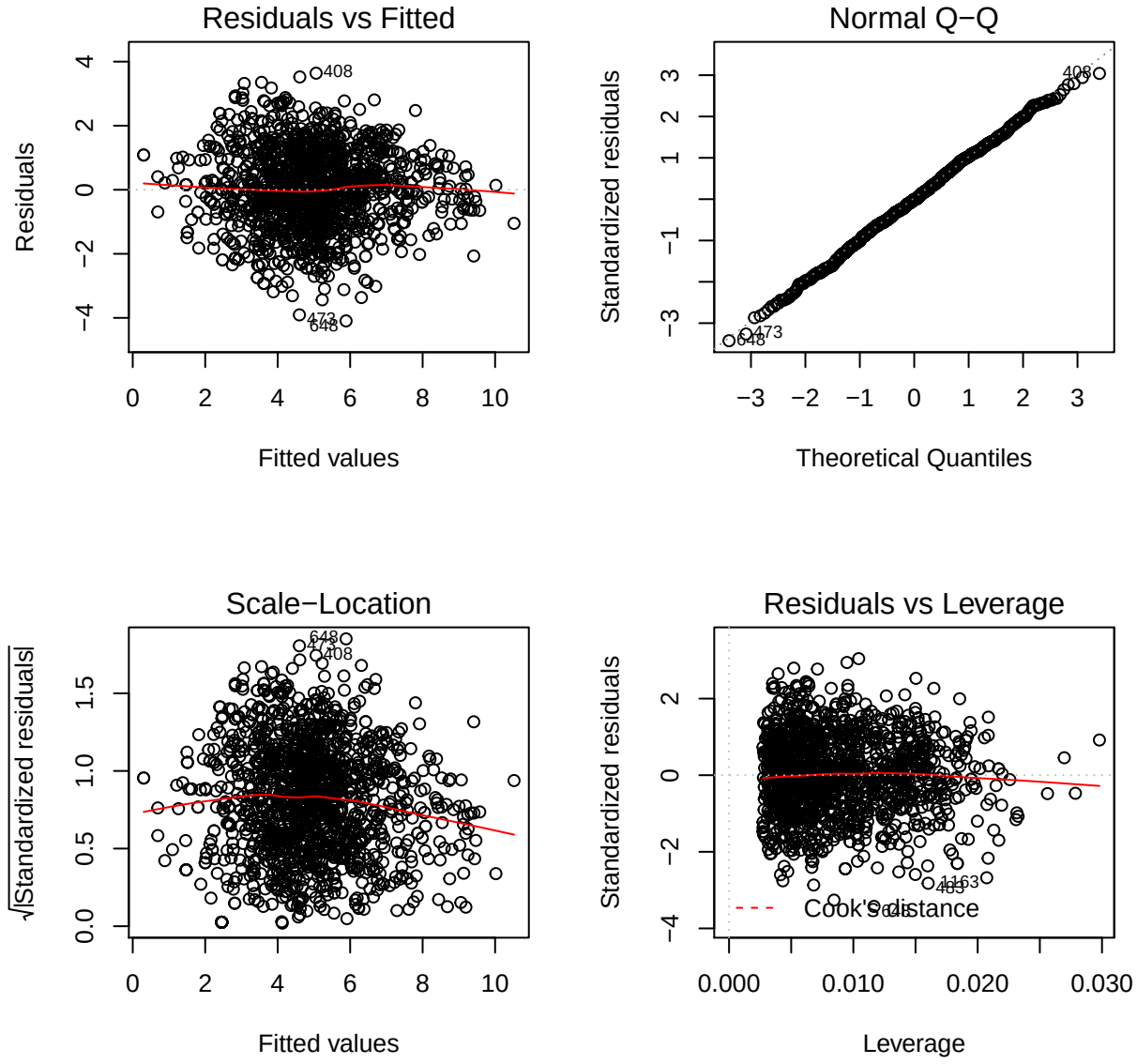


Comparing results between models, we see that RF does the best job with out of sample predictions, followed by the linear model and then BMA. These graphs visualize one random simulation to provide an idea of how the predictions look. The table on the graph shows the results when running the simulation n times. We see that while Random Forests produce slightly better RMSE than the linear regression model, the coverage is a little bit worse. Additionally, the Random Forest algorithm loses its interpretability by averaging many trees together. We want to present an interpretable model that does a good job with predictions while achieving proper coverage. Under this selection criteria, we determined that the linear regression model should be our final model. Now that we identified the model, we want to retrain the model on all of the training data.

Our final model can be summarised as follows:

$$\begin{aligned}\text{logprice} = \beta_0 + \beta_1 \text{year} + \beta_2 \text{dealer} + \beta_3 \text{expensive} + \beta_4 \text{authorstyle} + \beta_5 \text{endbuyer} \\ + \beta_6 \text{Interm} + \beta_7 \text{Surface} + \beta_8 \text{finished} + \beta_9 \text{Irgfont} + \beta_{10} \text{diff-origin} + \epsilon\end{aligned}$$

Residual Plot Analysis



Looking at the diagnostic plots, our final model seems to satisfy the assumptions of linear regression resonablly well. From the Residual vs Fitted plot we can see equally spread residuals around a horizontal line without any distinct patterns; The Normal Q-Q plot shows the residuals are almost normally-distributed. The Scale-Location plot shows that homoscedasticity is met. The Residual vs Leverage plot does not show any points that are influential or falls outside of Cook's distance line.

Summary Table

Table 4: Coefficient Summary for Final Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.000	12.576	-15.831	0.000	0.000	0.000
year	1.120	0.007	16.041	0.000	1.105	1.136
dealerLP	2.568	0.112	8.397	0.000	2.060	3.200
dealerR	5.356	0.103	16.260	0.000	4.374	6.558
expensive1	3.605	0.106	12.046	0.000	2.926	4.442
authorstyle1	0.388	0.148	-6.414	0.000	0.291	0.518
endbuyerD	0.788	0.101	-2.360	0.018	0.646	0.961
endbuyerEU	0.494	0.103	-6.874	0.000	0.404	0.604
Interm1	1.996	0.133	5.188	0.000	1.537	2.591
Surface	1.367	0.026	12.173	0.000	1.300	1.437
finished1	2.351	0.090	9.466	0.000	1.969	2.806
lrgfont1	2.436	0.120	7.442	0.000	1.926	3.080
diff_origin1	0.613	0.085	-5.788	0.000	0.519	0.724

Note:
The coefficients have been exponentiated.

Looking at the coefficient summary table, we can get the following conclusion.

First of all, in our model we got the most important variables *year*, *dealer*, *expensive*, *authorstyle*, *endbuyer* ,*Interm*, *Surface*, *finished*, *lrgfont*, *diff_origin*.

Then, holding all other variables constant,

- On average, one unit increase in *year* will lead to 12.0% increase in price. We are 95% confident the increase is between about 10.5% and 13.6%.
- In data manipulation part, we regrouped *dealer* by grouping 'L' and 'P' together as 'LP' since their corresponding average logprices are similar. Compared to dealer J, dealer LP is expected to lead to 156.8% increase in price, and we are 95% confident the decrease is between about 106.0% and 220.0%; Dealer R is expected to lead to 435.6% increase in price, and we are 95% confident the increase is between about 337.4% and 555.8%.
- If the painting is drawn by an artist whose paintings' average price is higher than $2000, the price is expected to increase by 260.5%. We are 95% confident the increase is between about 192.6% and 344.2%.
- If the authors name is introduced, the price is expected to decrease by 61.2%. We are 95% confident the decrease is between about 48.2% and 70.9%.

- In data manipulation part, we regrouped *endbuyer* by grouping 'U' (identity unknown), 'E' (expert organizing the sale) and blank data (no information) together as 'EU', grouping 'B' (buyer) and 'C' (collector) together as 'C' and left 'D' (dealer) as it was. Compared to an endbuyer in group 'C', a group 'D' endbuyer is expected to lead to 21.2% decrease in price, and we are 95% confident the decrease is between about 4.1% and 35.4%; A group 'EU' endbuyer is expected to lead to 50.6% decrease in price, and we are 95% confident the decrease is between about 39.6% and 59.6%.
- If an intermediary is involved in the transaction, the price is expected to increase by 99.6%. We are 95% confident the increase is between about 53.7% and 159.1%.
- We expect 10% increase in *Surface* will increase the price by 3.0% ($1.1^{\hat{\beta}_1} - 1$). We are 95% confident the increase is between about 2.5% and 3.5%.
- If the the painting is finished, the price is expected to increase by 135.1%. We are 95% confident the increase is between about 96.9% and 180.6%.
- If the dealer devotes an additional paragraph, the price is expected to increase by 143.6%. We are 95% confident the increase is between about 92.6% and 208.0%.
- If origin_author is different than origin_cat, the price is expected to decrease by 38.7%. We are 95% confident the increase is between about 27.6% and 48.1%.

Looking at the p-values, we find that all variables are extremelly important.

Prediction Intervals

We used the built-in function for linear model to obtain the prediction intervals, which are usually wider than confidence interval due to the fact that they also take consideration of variance that comes from the true error term. In fact, one of the reasons that we finally decided to use linear model is that its prediction interval is easier to obtain than those of Lasso and trees, which require us to manually use other methods to estimate the intervals for predictions.

Assessment of the final model

Model Evaluation

Our assessment of the final model began with running a simulation analysis on the out-of-sample performance of the model. We thought that this check would give us similar result to our results on the leaderboard. The results from the simulation show that we expect our model to have an RMSE of approximatly 1600 with 95% coverage. When we submitted the predictions on the test data, we saw that our results were consistent with our simulations: an RMSE of 1600 and 95% coverage.

Our goal for improving our Part I model was to keep the same coverage and lower RMSE through inducing bias and reducing variance. We achieved this goal and lowered the RMSE by approximatly 33% (from 2360 to 1600).

We did one last check to ensure that we were meeting all the assumptions in linear regression. There could be a multicollinearity problem when putting multiple predictors into a regression.

Table 5: VIF of final model

	GVIF	Df	GVIF^(1/(2*Df))
year	1.345	1	1.160
dealer	1.779	2	1.155
expensive	1.120	1	1.058
authorstyle	1.181	1	1.087
endbuyer	1.982	2	1.187
Interm	1.606	1	1.267
Surface	1.082	1	1.040
finished	1.129	1	1.062
lrgfont	1.336	1	1.156
diff_origin	1.430	1	1.196

To check for the multicollinearity problem, we used variance inflation factor (VIF). The result is in the table above. The issue of multicollinearity is negligible since no VIF exceeds 5.

Model Testing

See model selection. Here we trained the model and tested on unseen data 1,000 times to ensure stable results. The leaderboard results aligned with these findings.

Predictions of Validation set and Top 10 paintings

Table 6: Top 10 paintings

fitted	year	dealer	expensive	authorstyle	endbuyer	Interm	Surface	finished	lrgfont	diff_origin
15072.791	1776	R	1	0	C	1	5.690	1	1	0
12213.427	1769	R	1	0	C	1	7.560	1	1	0
10452.386	1767	R	1	0	C	1	7.788	1	1	0
9821.674	1777	R	1	0	C	1	6.692	0	1	0
9770.848	1776	R	1	0	C	1	7.039	0	1	0
9199.615	1769	R	1	0	C	1	6.653	1	1	0
8292.477	1767	R	1	0	C	1	8.613	1	1	1
7905.600	1777	R	0	0	C	1	7.366	1	1	0
6489.631	1777	R	1	0	D	0	5.606	1	1	0
6286.282	1777	R	0	0	C	1	6.633	1	1	0

Using our model for predicting price for validation data set, we got our top 10 valuable paintings.From this we can learn what are some desirable features of the paintings based on our model through observing these valuable paintings all share certain common features, such as they are all from the same dealer, R. In addition, endbuyers are mostly from category C, the dealer devotes an additional paragraph and an intermediary is involved in the transaction etc. This is quite expected due to the way we constructed our model.

Conclusion

In EDA process of part-I, we imputed the NA's in *Surface* to median value of *Surface*, we discarded the variable *authorstandard* and we grouped *dealer* and *endbuyer* into 4 and 5 levels respectively. And then we fitted a linear model. However, in part-II, our model were not good as we expected. We returned to EDA process and realized that there was still some infomation in raw data we did not pay enough attention. And thus we reworked the data manipulation process. In part-II, we devided the surface data into 8 groups based on the correlated data and imputed median value for each group respectively. We created a binary variable *expensive* representing if the painting is drawn by a painter whose paintings' averaging price is higher than $2000. And we also regrouped dealer and endbuyer to less levels to avoid overfitting. After data re-manipulation, the model performed better.

Before the test data was updated, we always felt upset because the coverage was always lower than our expectaiton even if we thought our model was already good enough. Some of our team members were so disappointed that they even thought about discarding *year* since this operation improved the coverage a lot. But luckily we were pretty sure that all operations need to be reasonable. We could not randomly discard some variables just for better coverage and RMSE. So we still followed the rigorous modeling process to fit our model, even if we did not get a score in wercker.

After a few attempts, we had to give up our most prefered model Random Forests since it could not yield prediction interval and the quantile method was not ideal. We then used Bayesian Model Averaging to do feature selection and cross referenced the results of RF and BMA and fitted a new linear model, which achieved relatively great performance. We found that succinct model usually have higher coverage and easy interpretability so it's easier for clients to understand and linear model is generally a great model choice.

In terms of painting price prediction, we think that due to the particularity of art, the author often has a great impact on the price. Besides, year of sale, dealer, if the authors name is introduced, type of end buyer, whether or not an intermediary is involved in the transaction, surface of painting, if the painting is finished, if the dealer devotes an additional paragraph, if origin of author is different than origin of painting also have a great impact on the price of paintings.

What we learned

A key takeaway from this project is to trust our instincts as statisticians. We knew that something could not be right when we achieved good results from testing our model on cross-validated data but could not beat the null model on the leaderboard. Instead of doubting our results because the leaderboard said we were not doing well, we went with our gut instinct that something must be wrong in the leaderboard calculations and to continue building our model using cross validation as our out-of-sample performance checks. This is a valuable lesson to not let the stakes of the situation (a final exam grade in this case) get in the way of conducting a proper analysis.

If we had more time...

We would have gone back and checked our scores for the various other models we tested along the way. Further data cleaning and variable manipulation could also continue to help with predictions.

Reference

[1] Hoeting, Jennifer A., et al. “Bayesian Model Averaging: A Tutorial.” Statistical Science, vol. 14, no. 4, 1999, pp. 382–401. JSTOR, www.jstor.org/stable/2676803.

Appendix

Table 7: Coefficient Summary for Full Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.000000e+00	113.600	-2.156	0.031	0.000000e+00	0.000000e+00
year	1.148000e+00	0.064	2.147	0.032	1.012000e+00	1.302000e+00
dealerLP	2.307320e+77	81.821	2.177	0.030	4.428257e+07	1.202217e+147
dealerR	1.658112e+06	56.414	0.254	0.800	0.000000e+00	1.941402e+54
origin_catF	0.000000e+00	42.303	-1.472	0.141	0.000000e+00	1.002949e+09
origin_catI	0.000000e+00	53.005	-0.163	0.870	0.000000e+00	2.562458e+41
origin_catO	1.590247e+16	87.371	0.427	0.669	0.000000e+00	4.442916e+90
diff_origin1	3.510414e+06	66.989	0.225	0.822	0.000000e+00	4.208847e+63
authorstyle1	0.000000e+00	71.457	-2.328	0.020	0.000000e+00	0.000000e+00
endbuyerD	2.526748e+31	47.486	1.523	0.128	0.000000e+00	7.312775e+71
endbuyerEU	0.000000e+00	47.972	-0.939	0.348	0.000000e+00	2.074942e+21
Interm1	0.000000e+00	68.049	-0.316	0.752	0.000000e+00	4.300218e+48
Surface	2.098000e+00	14.866	0.050	0.960	0.000000e+00	9.745311e+12
materialCatcopper	2.000000e-03	63.797	-0.095	0.925	0.000000e+00	5.472076e+51
materialCatother	1.351184e+13	43.090	0.702	0.483	0.000000e+00	7.018161e+49
materialCatwood	1.262517e+07	42.607	0.384	0.701	0.000000e+00	2.540269e+43
nfigures1	3.395525e+06	83.926	0.179	0.858	0.000000e+00	1.099276e+78
engraved1	8.031108e+11	92.138	0.298	0.766	0.000000e+00	2.586572e+90
prevcoll1	1.099501e+26	101.895	0.588	0.556	0.000000e+00	7.287982e+112
paired1	1.689798e+10	31.174	0.755	0.450	0.000000e+00	6.172932e+36
finished1	1.110000e-01	50.211	-0.044	0.965	0.000000e+00	6.733914e+41
lrgfont1	3.949457e+18	67.323	0.636	0.525	0.000000e+00	9.129766e+75
lands_sc1	1.900000e-02	54.191	-0.073	0.941	0.000000e+00	2.791364e+44
lands_elem1	0.000000e+00	34.633	-0.397	0.691	0.000000e+00	3.455549e+23
othgenre1	0.000000e+00	64.604	-1.670	0.095	0.000000e+00	1.562146e+08
discauth1	1.176828e+209	136.877	3.517	0.000	2.771968e+92	Inf
year:dealerLP	9.060000e-01	0.046	-2.139	0.033	8.280000e-01	9.920000e-01
year:dealerR	9.940000e-01	0.032	-0.186	0.853	9.340000e-01	1.058000e+00
year:origin_catF	1.035000e+00	0.024	1.451	0.147	9.880000e-01	1.085000e+00
year:origin_catI	1.004000e+00	0.030	0.140	0.889	9.470000e-01	1.065000e+00
year:origin_catO	9.810000e-01	0.049	-0.395	0.693	8.900000e-01	1.080000e+00
year:diff_origin1	9.910000e-01	0.038	-0.227	0.821	9.200000e-01	1.068000e+00
year:authorstyle1	1.099000e+00	0.040	2.334	0.020	1.015000e+00	1.189000e+00
year:endbuyerD	9.610000e-01	0.027	-1.502	0.133	9.110000e-01	1.012000e+00
year:endbuyerEU	1.026000e+00	0.027	0.948	0.343	9.730000e-01	1.082000e+00
year:Interm1	1.012000e+00	0.038	0.299	0.765	9.380000e-01	1.090000e+00
year:Surface	1.000000e+00	0.008	0.002	0.999	9.840000e-01	1.017000e+00
year:materialCatcopper	1.005000e+00	0.036	0.125	0.900	9.360000e-01	1.078000e+00
year:materialCatother	9.830000e-01	0.024	-0.693	0.489	9.380000e-01	1.031000e+00
year:materialCatwood	9.910000e-01	0.024	-0.357	0.721	9.460000e-01	1.039000e+00
year:nfigures1	9.920000e-01	0.047	-0.166	0.868	9.040000e-01	1.089000e+00
year:engraved1	9.840000e-01	0.052	-0.301	0.764	8.890000e-01	1.091000e+00
year:prevcoll1	9.670000e-01	0.057	-0.591	0.555	8.640000e-01	1.082000e+00
year:paired1	9.870000e-01	0.018	-0.763	0.446	9.530000e-01	1.021000e+00
year:finished1	1.002000e+00	0.028	0.064	0.949	9.480000e-01	1.059000e+00
year:lrgfont1	9.770000e-01	0.038	-0.610	0.542	9.070000e-01	1.053000e+00
year:lands_sc1	1.003000e+00	0.031	0.088	0.930	9.440000e-01	1.065000e+00
year:lands_elem1	1.008000e+00	0.019	0.418	0.676	9.700000e-01	1.047000e+00
year:othgenre1	1.063000e+00	0.036	1.681	0.093	9.900000e-01	1.142000e+00
year:discauth1	7.620000e-01	0.077	-3.518	0.000	6.550000e-01	8.870000e-01