Part-II Complex Model

Chenxi Wu, George Lindner, Qianyin Lu, Yi Mi

11/30/2019

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.

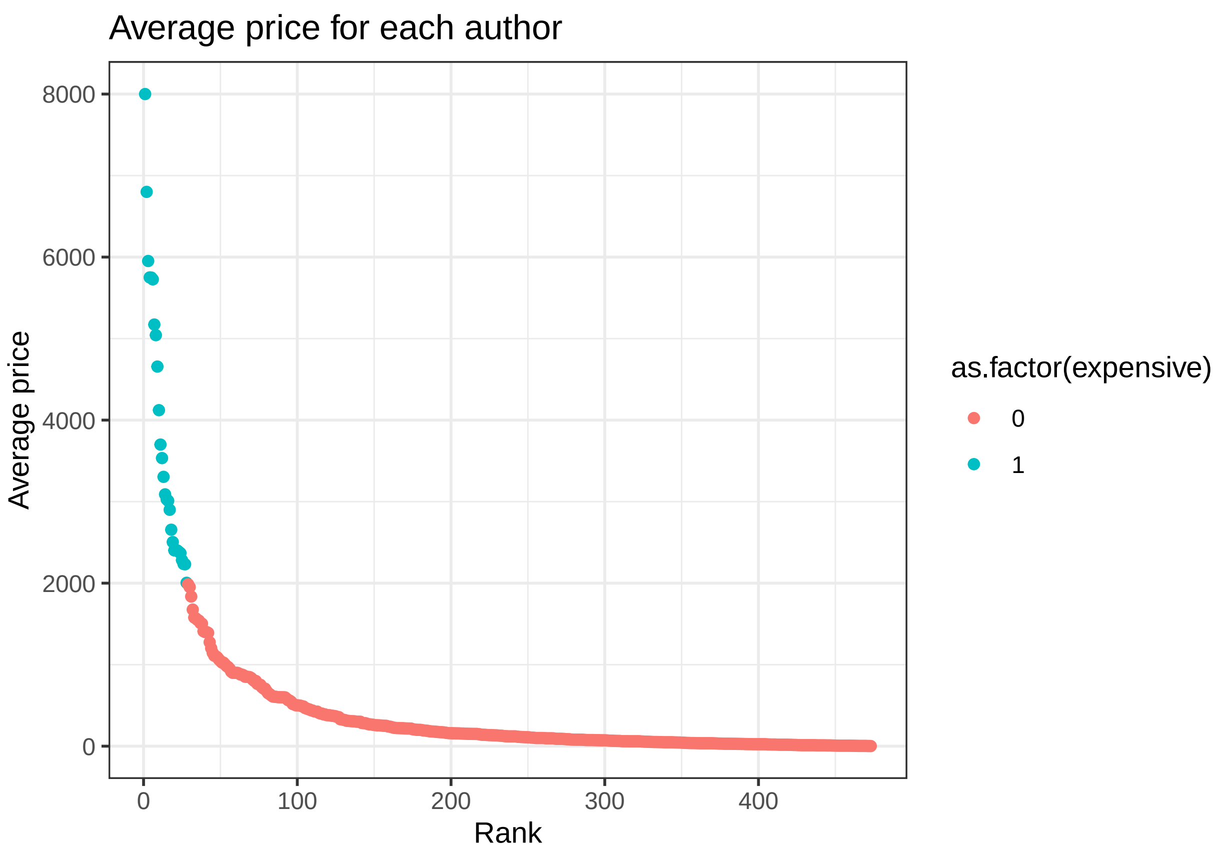
2) To determine which paintings were overpriced or and which were underpriced.
   The first objective could be accomplished through EDA and modeling. Getting to know the dataset through EDA helps our team identify relationships in the data and develop a sense of which variables might be important for 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.

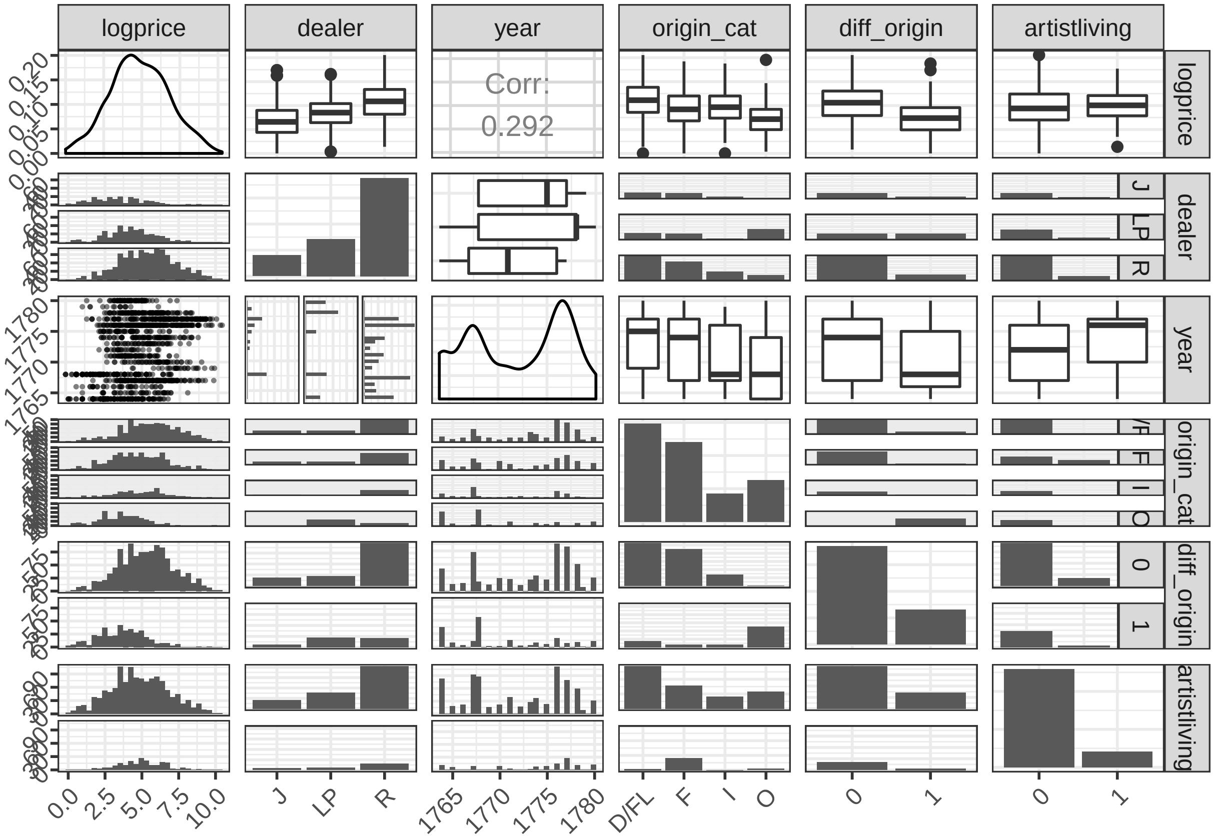
EDA and Data Manipulation

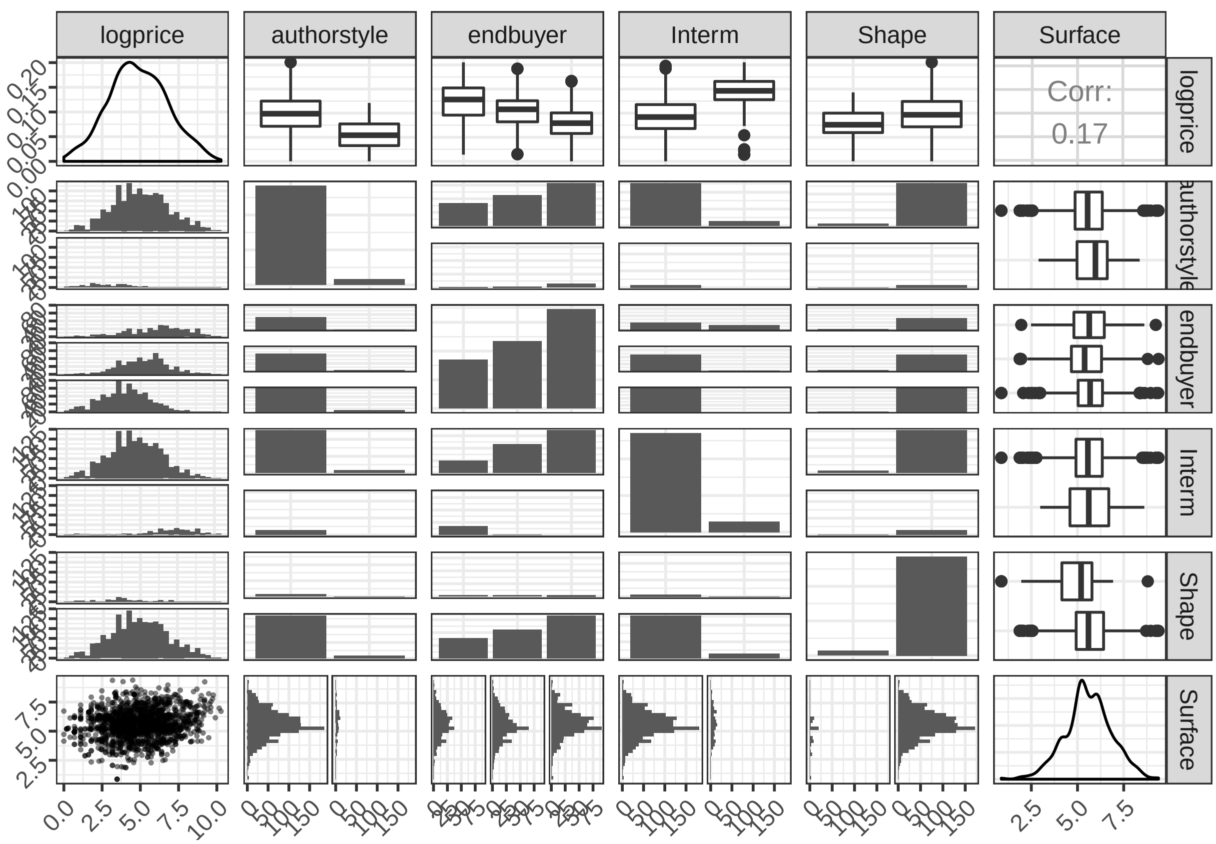
Based on EDA of Part-I, we improved our data manipulation on our data as follows:

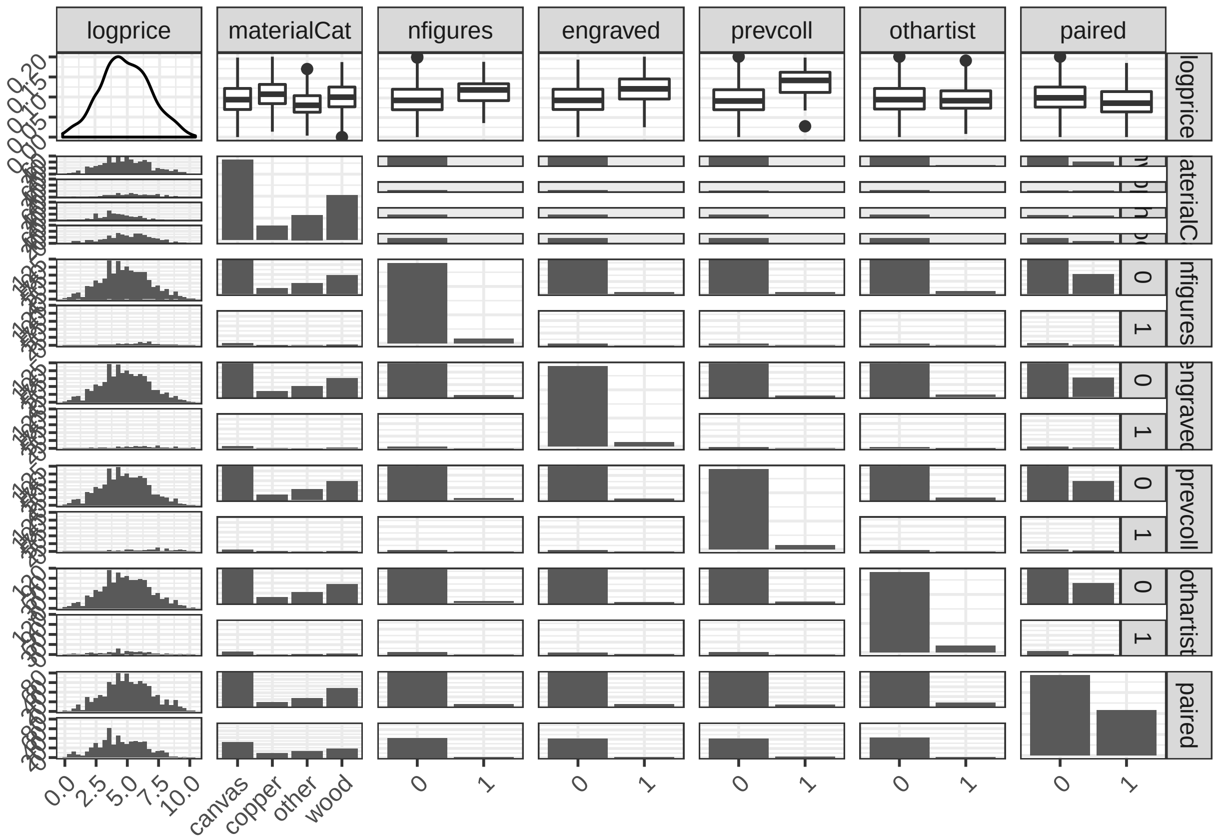
- *position* has values greater than 1 which should be data entry errors, we divided them by 100 to get the right value.
- The original dataset contains lots of missing values and NA's, like *winningbiddertype*, *endbuyer*, *authorstyle*, *Interm* and *type_intermed*, we filled the missing values with "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 sginificant 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*.




logprice
dealer
year
origin_cat
diff_origin
artistliving
Corr:
0.292
J
LP
R
D/FL
F
I
O
0
1
0.0
2.5
5.0
7.5
10.0
1765
1770
1775
1780


logprice
authorstyle
endbuyer
Interm
Shape
Surface
Corr:
0.17
0.0
2.5
5.0
7.5
10.0


logprice
materialCat
nfigures
engraved
prevcoll
othartist
paired
0.00
0.05
0.10
0.15
0.20
0
1
0.0
2.5
5.0
7.5
10.0
canvas
copper
other
wood

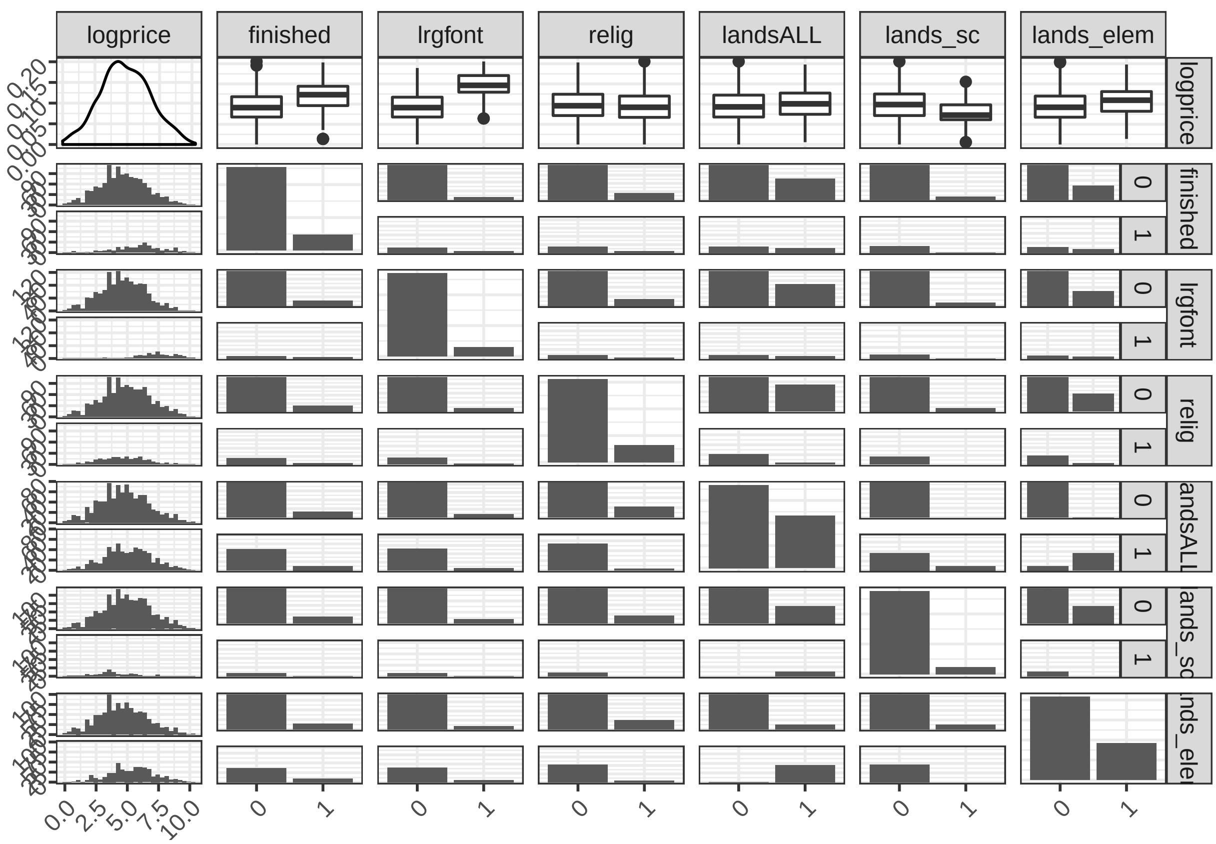
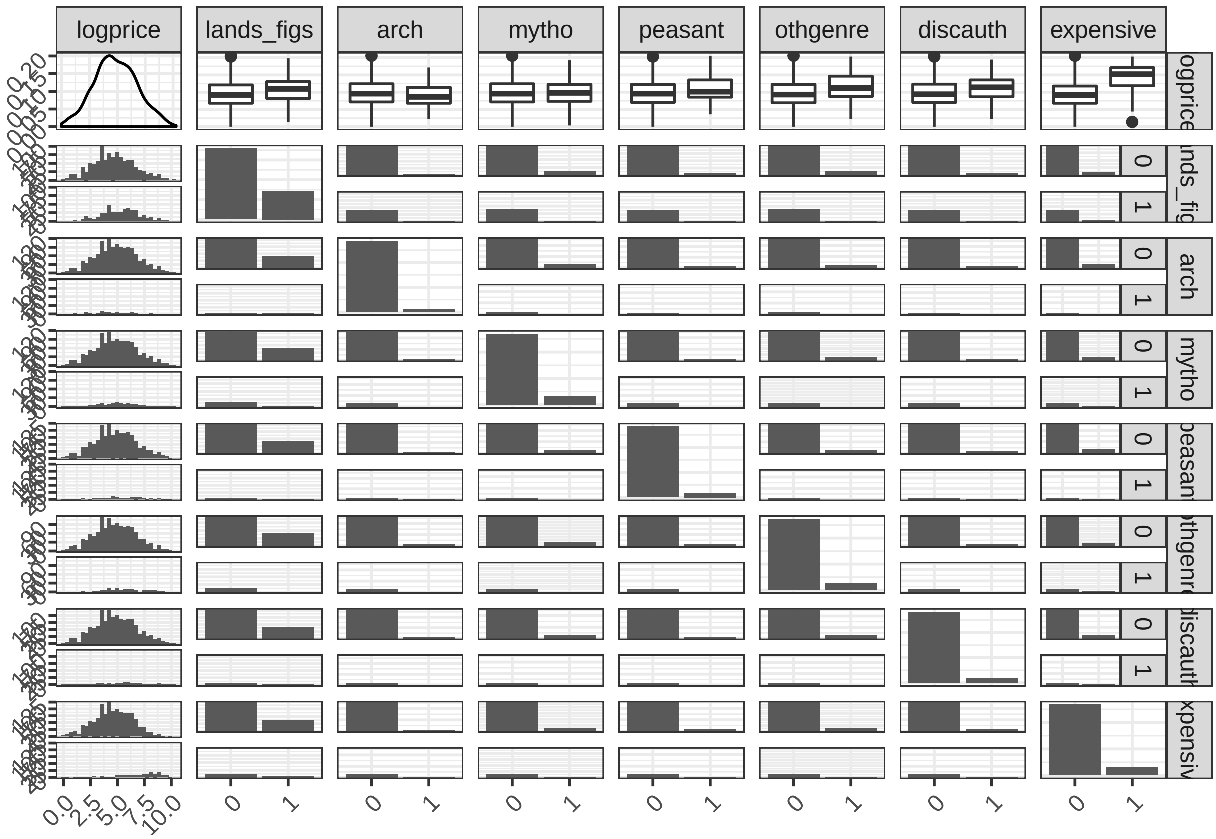

logprice
finished
lrgfont
relig
landsALL
lands_sc
lands_elem
0.00
0.05
0.10
0.15
0.20
0
1
0.0
2.5
5.0
7.5
10.0

Table 1: Results from Preliminary Model

Bias	Coverage	MaxDeviation	MeanAbsDeviation	RMSE
120.29	95.6	52516.79	551.74	2363.27



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.

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 DEA above. Below is the important variable plot and the top 10 most important variable table. The 10 most important variables are *experience*, *year*, *Surface*, *endbuyer*, *dealer*, *materialCat*, *origin_cat*, *paired*, *Irgfont* and *finished*. From random forrest we obtained the 5 least important variables and discarded them in further modeling. The variables are displayed in the table below.

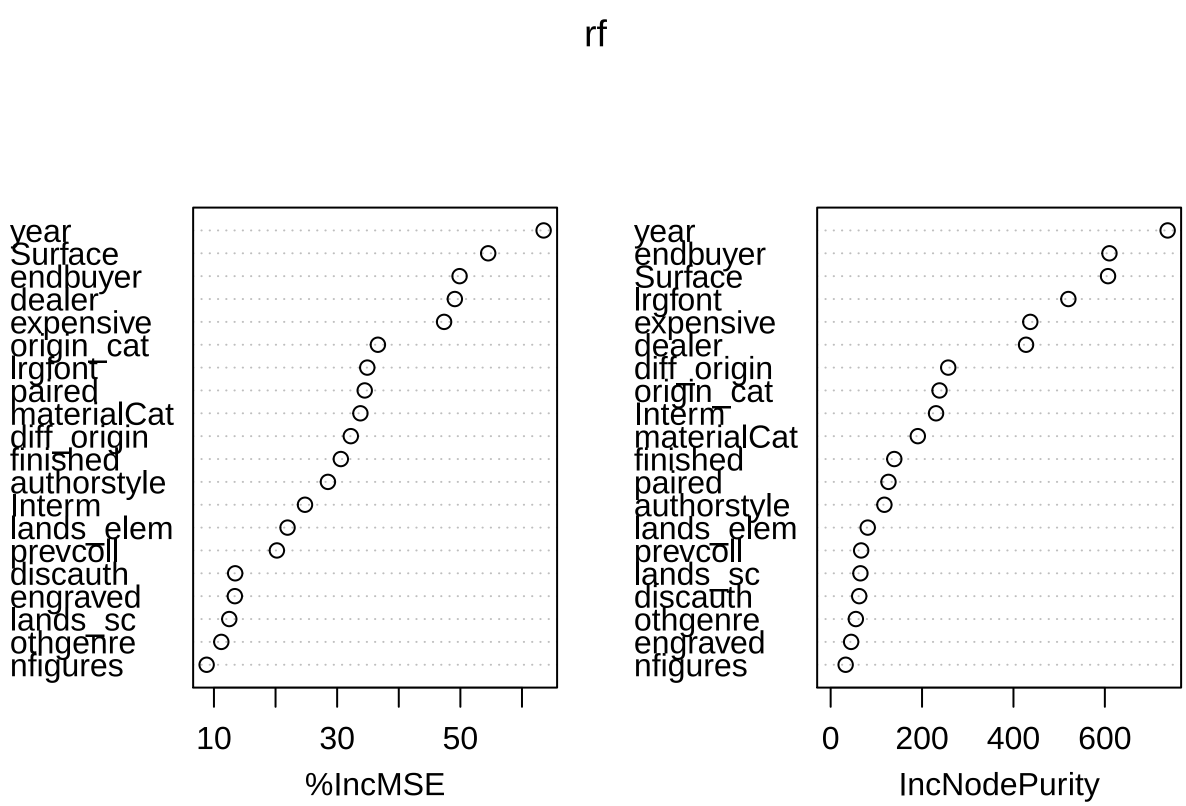


Table 2: Least 5 important variables of RF

	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 the narrower interval. So we move on to other variable selection method like Bayesian Model Averaging.

still need predictive checks

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:

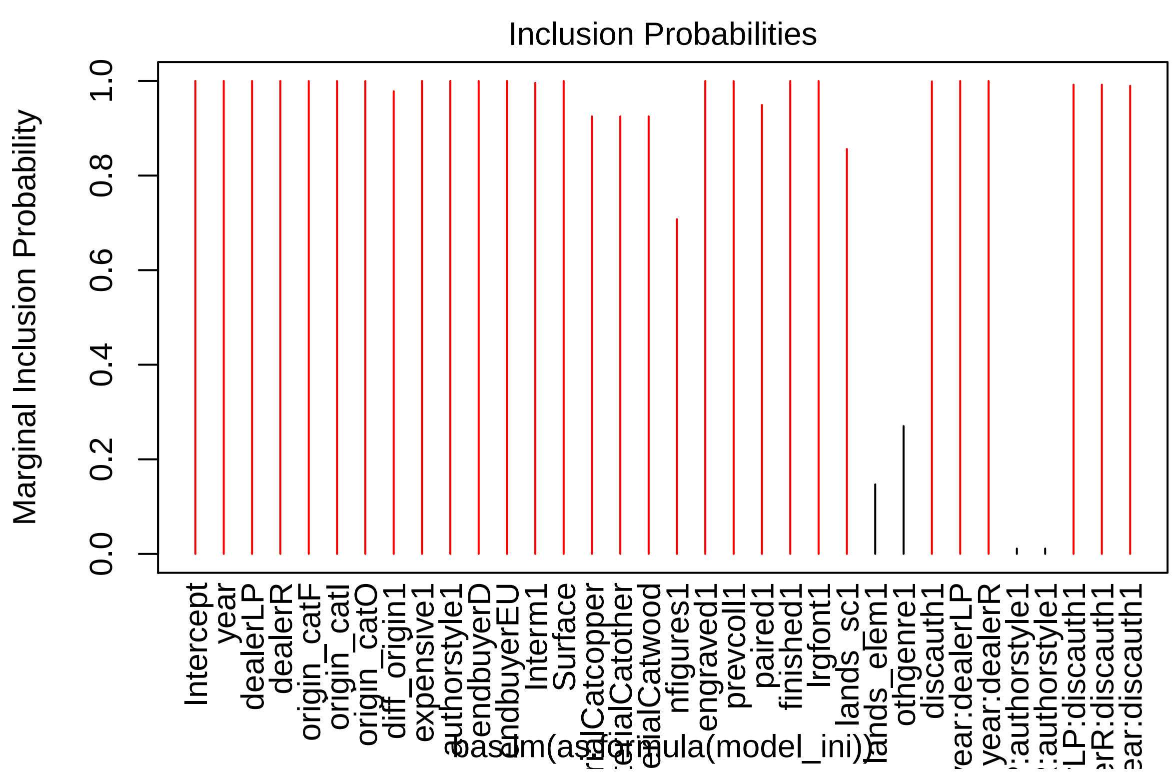
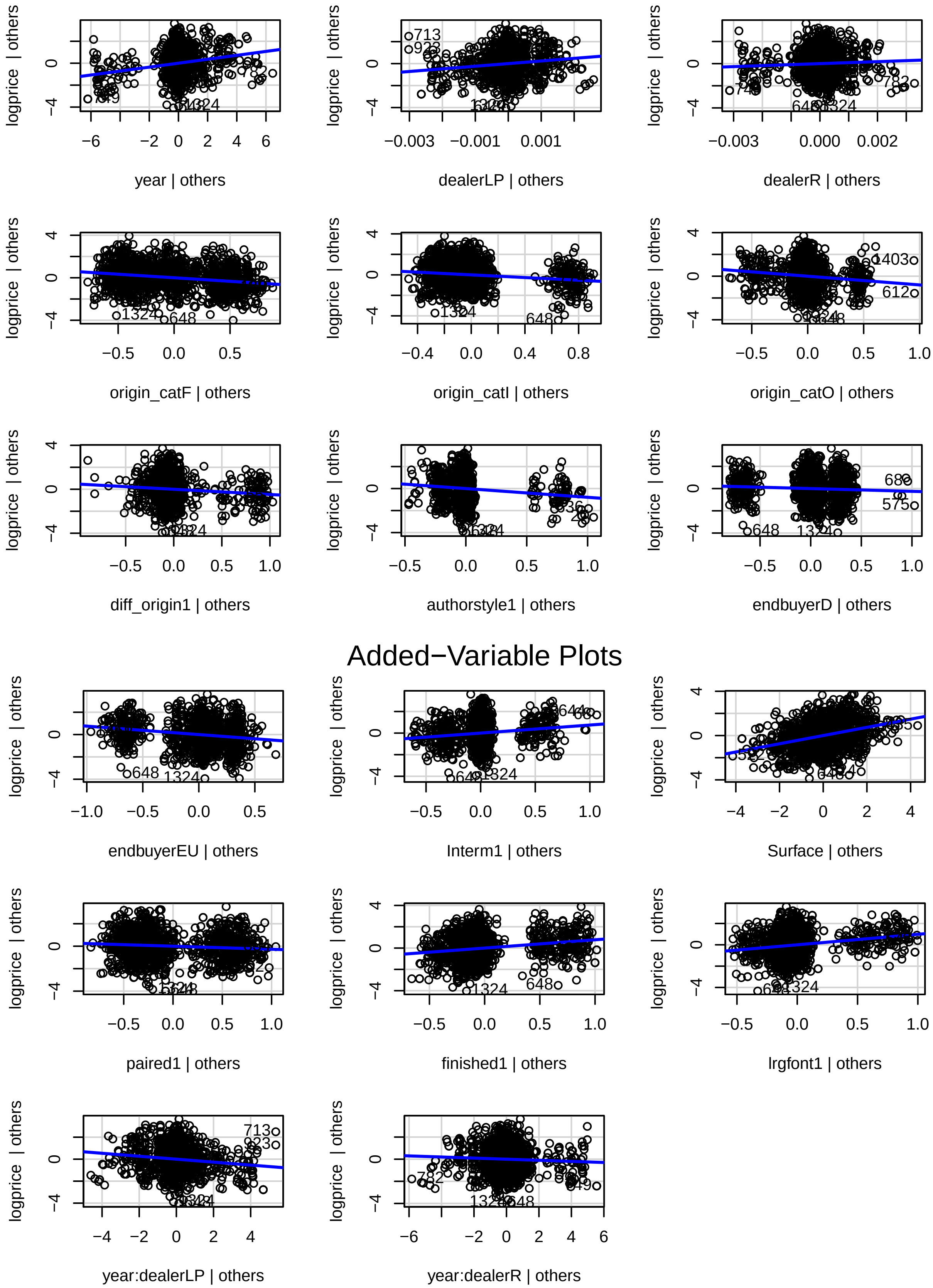


Table 3: Coefs and C.I. of best BMA model

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

From the marginal inclusion probability plot, we should exclude *materialCat*, *nfigures*, *lands_elem*, *auhtorstyle*:*dealer* since their margnial inclusion probability is less than 0.5.

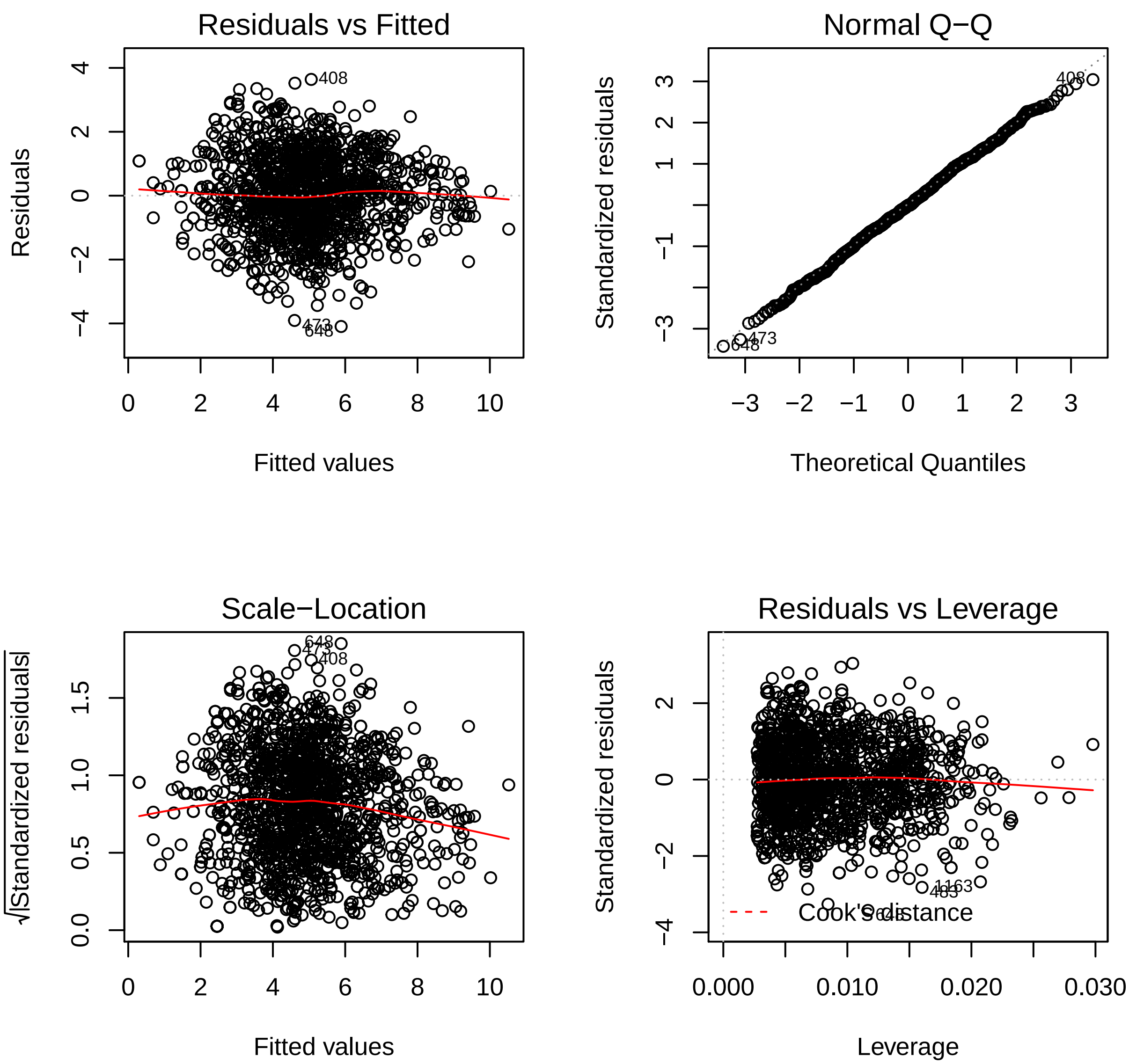
After cross referencing the result of random forrest and BMA, we decided to give another shot with simple models. So we discarded the variables that are not significant and refit a linear model. The problem of overfitting still exists. To further select variables, we used Added variable plots, which shows 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. Our final model can be summarised as follows:

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

Assessment of the final model



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.

Table 4: 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. Issue of multicollinearity is negligible for no VIF exceeds 5.

Predictions of Validation set and Top 10 paintings

Table 5: 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

Table 6: 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

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