STA 521 - Final Project Part I

FP-Team 01: Qianyin Lu, George Lindner, Chenxi Wu, Yi Mi

December 7th, 2019

1. Introduction: Summary of problem and objectives

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. These variables were all possible predictor variables in modeling the response variable, the selling price of a 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.

After arriving at a final model, we are able to answer these primary questions. Any variables that appear in the model will be important in driving painting prices, and observing residuals will enable us to determine if a painting was over or underpriced.

We had 1,500 observations to train the model on, along with 750 observations held out as a testing set. There were a total of 59 variables in the dataset, both categorical and continuous.

2. Exploratory Data Analysis:

Initial Data Cleaning

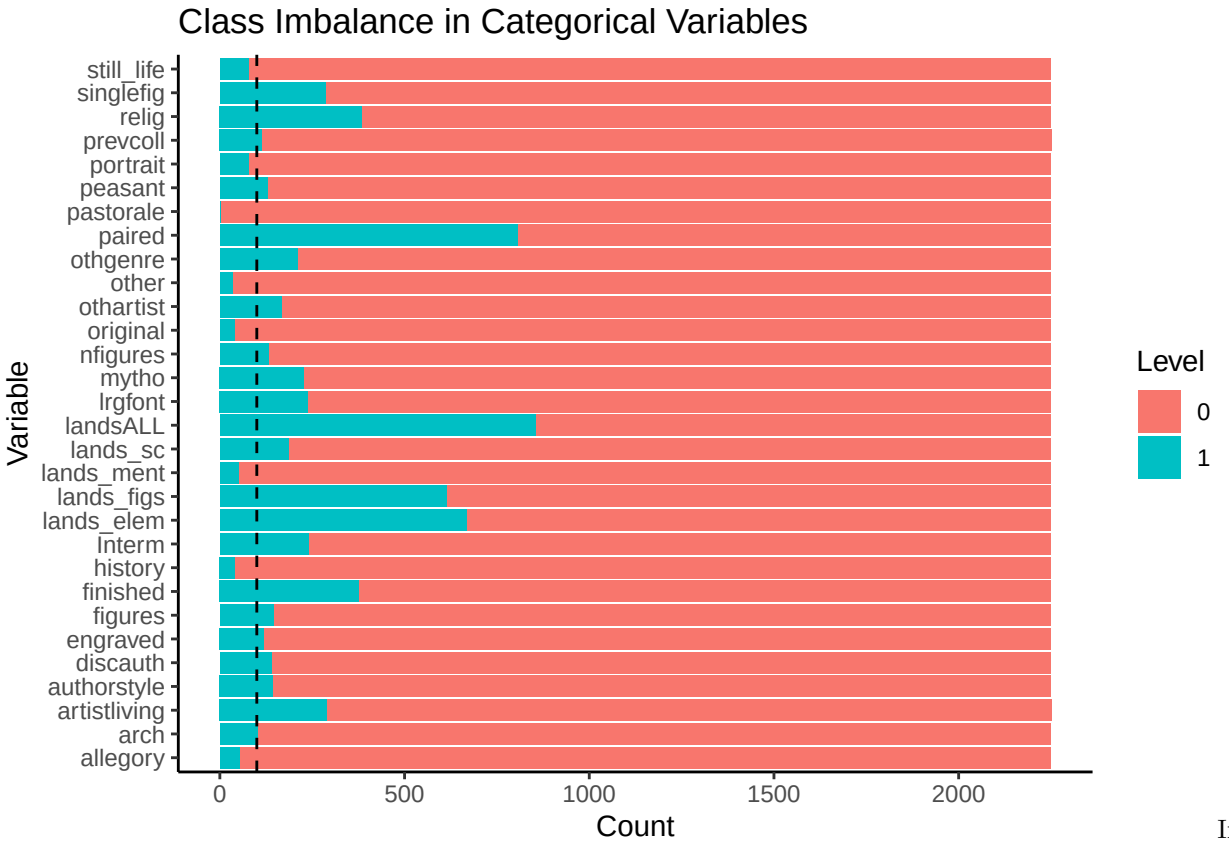
We began our data cleaning process by reading the codebook for a better understanding of what each variable in the data represented. Several predictors in the dataset were redundant and therefore removed to avoid high correlation among the predictors. Examples of this include the variable *sale*, which 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.

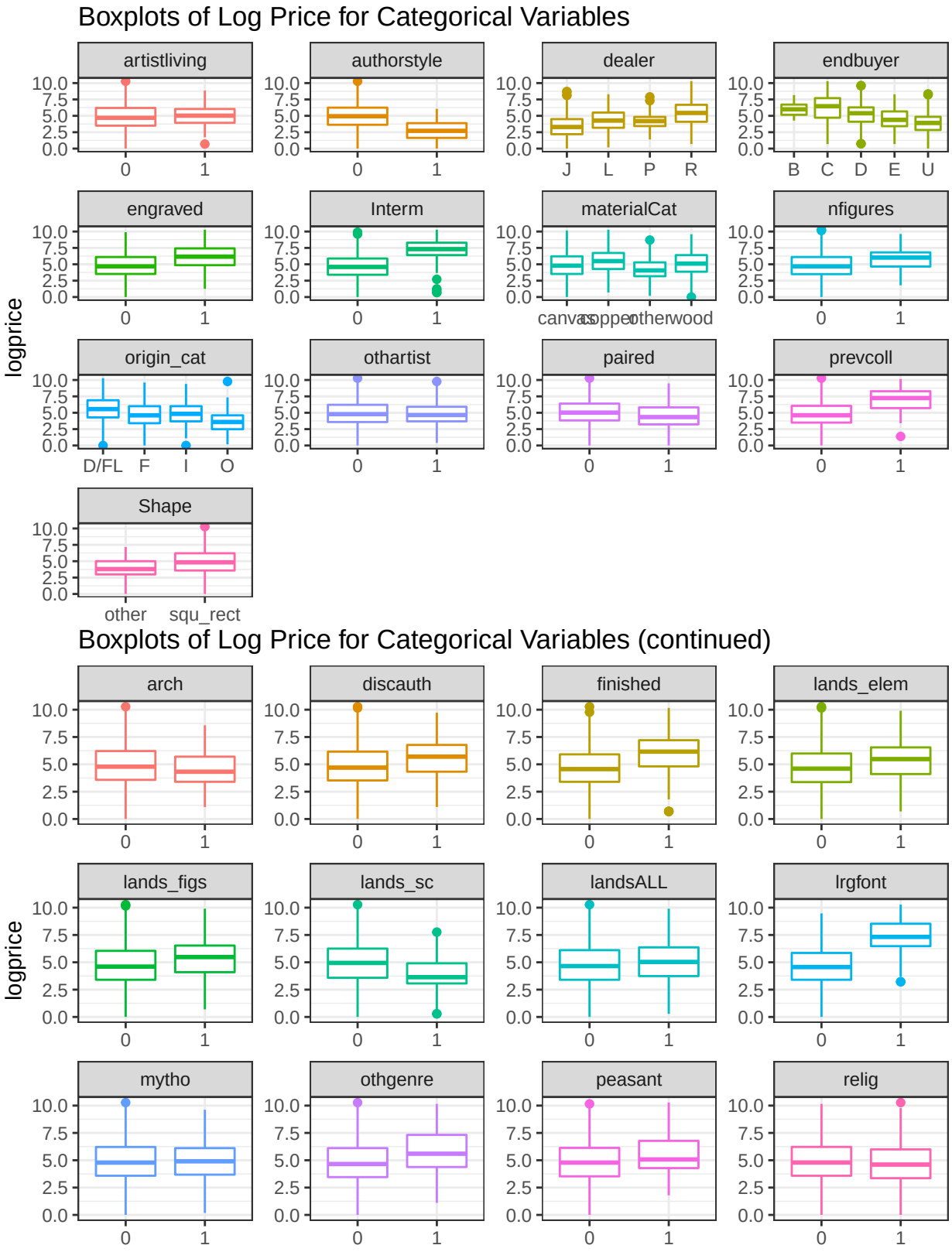
Plot 1



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



The boxplots above help us identify which variables could be important in predicting a painting's price. They

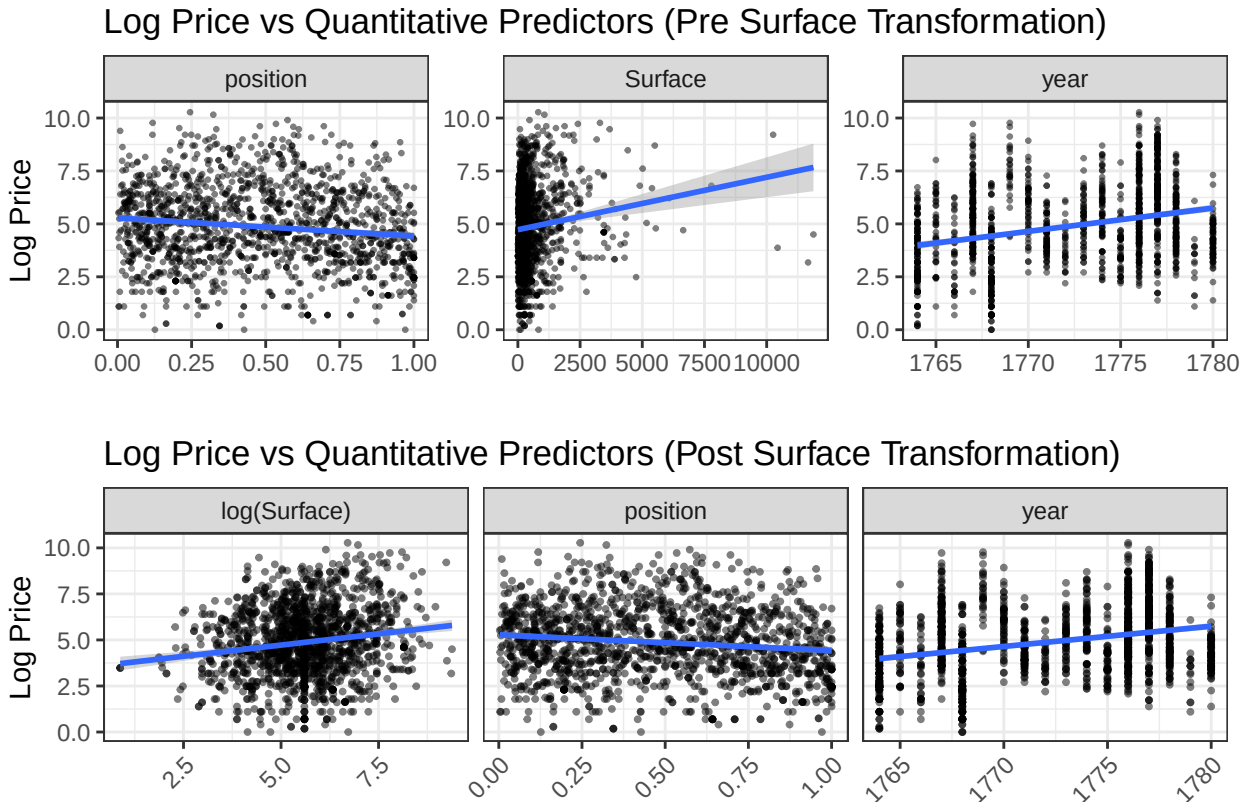
also help us in our variable selection process by displaying variables that have similar prices in all of their categories. After inspecting the boxplots, we determined that *mytho*, *landsALL*, *relig*, and *othartist* were not useful for prediction. Variables that may be important include, but are 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.

After considering our EDA plots, we determined the 10 variables that we thought would be most useful in predicting *logprice*.

With the data cleaned and important variables identified, we move to the next step of the process: modeling the data.

Table 1: 10 Most Important Predictor Variables, from EDA

Rank	Variable
1	log(Surface)
2	lrgfont
3	Interm
4	authorstyle
5	prevcoll
6	origin_cat
7	engraved
8	finished
9	discauth
10	dealer

3. Development and Assessment of Initial Model:

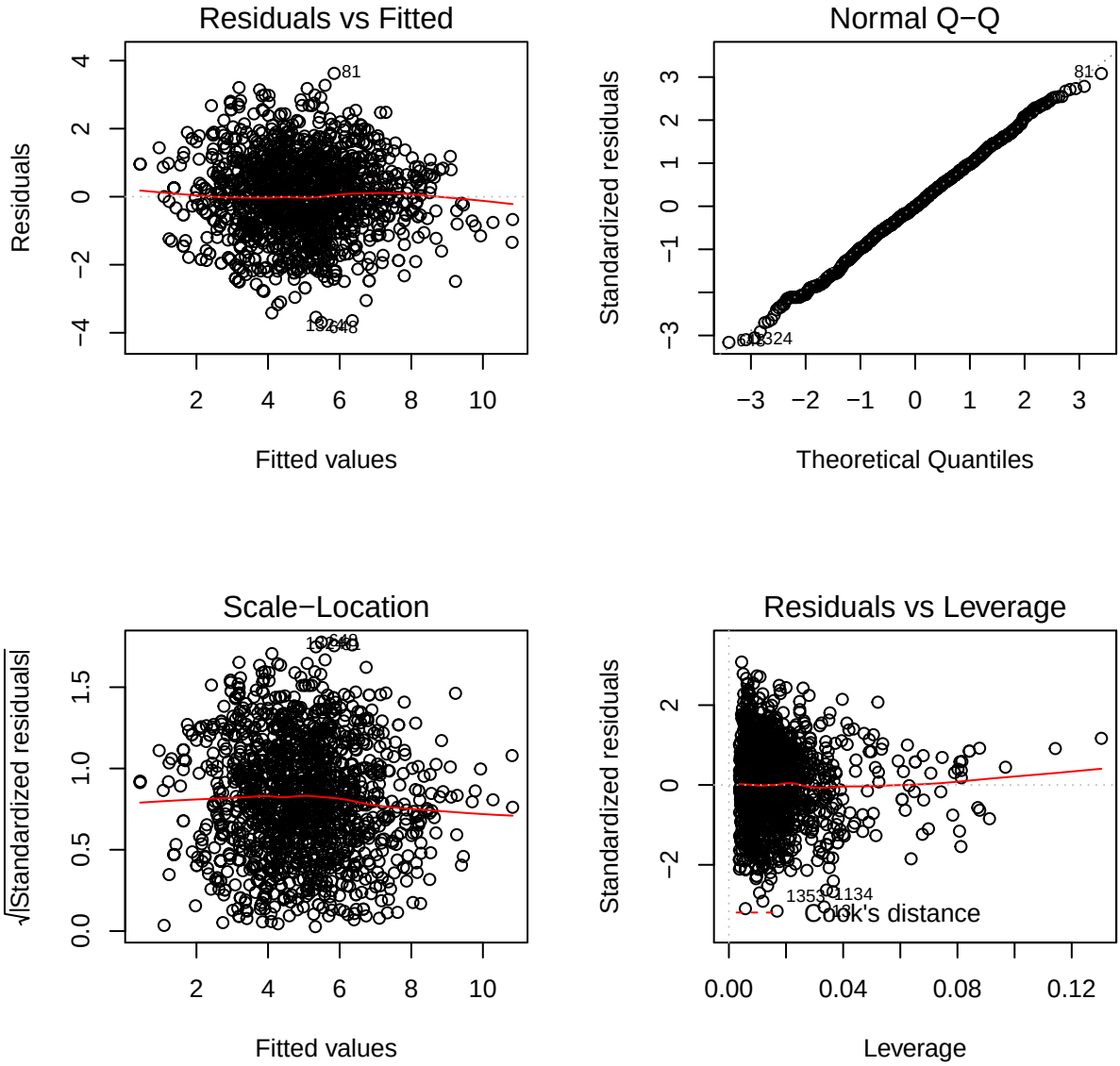
Model Development

We considered three additional variables in addition to the ten most important variables identified through EDA. These variables were *lands_sc*, *endbuyer*, and *year*. We chose to include these three variables because the EDA suggested that they might add some predictive power to our model. Approximately 62.2% of the variation in *logprice* can be explained by the predictors in our initial model, according to the summary output below. Next, we used step-wise variable seleciton with AIC as our criteria to ensure that each variable reduced RSS enough to justify including the variable in the model. Step-wise AIC selection returned our full model, indicating that we did a good job selecting predictor variables through the EDA.

Next, we considered interaction terms for our predictor variables. We again used intuition as our method for introducing interactions in the model. Interactions that we considered were *authorstyle* with *log(Surface)*, *Interm* with *log(Surface)*, and *discauth* with *log(Surface)*. We chose the first term because both the style of a painting and the size of the painting could be important in determining the price. Various styles of paintings might increase in value at different rates as the size of the painting changes. The second interaction term considers the dealer engaging with the authenticity of the painting and the size of the painting. Authentic paintings could increase in value as the size increases at a different rate than non-authentic paintings. We also thought that whether an intermediary was involved could be an important interaction with the size of the painting.

We fit our initial model on these predictors and interaction terms. Included below are our model plots and the summary of the model.

Model Plots



Model Summary

```
##
## Call:
## lm(formula = logprice ~ log(Surface) + lrgfont + Interm + authorstyle +
##     prevcoll + origin_cat + engraved + finished + discauth +
##     dealer + lands_sc + endbuyer + year + authorstyle:log(Surface) +
##     Interm:log(Surface) + log(Surface):discauth, data = paint_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6963 -0.7402 -0.0144  0.7876  3.6259
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.210e+02  1.290e+01 -17.130  < 2e-16 ***
```

```
## log(Surface)                3.530e-01  2.871e-02  12.294  < 2e-16 ***
## lrgfont1                    8.354e-01  1.207e-01   6.919 6.76e-12 ***
## Interm1                     7.832e-02  4.722e-01   0.166 0.868284
## authorstyle1               -7.890e-01  6.314e-01  -1.250 0.211633
## prevcoll1                   8.159e-01  1.426e-01   5.720 1.29e-08 ***
## origin_catF                -6.358e-01  7.643e-02  -8.318  < 2e-16 ***
## origin_catI                -7.276e-01  1.064e-01  -6.838 1.17e-11 ***
## origin_catO                -8.452e-01  1.103e-01  -7.665 3.23e-14 ***
## engraved1                   7.141e-01  1.438e-01   4.964 7.70e-07 ***
## finished1                   8.011e-01  9.086e-02   8.817  < 2e-16 ***
## discauth1                   1.053e+00  6.603e-01   1.595 0.110873
## dealerL                     1.298e+00  1.287e-01  10.083  < 2e-16 ***
## dealerP                     3.154e-01  1.586e-01   1.988 0.046973 *
## dealerR                     1.791e+00  1.032e-01  17.357  < 2e-16 ***
## lands_sc1                  -4.232e-01  1.161e-01  -3.645 0.000276 ***
## endbuyerC                  -2.950e-01  3.254e-01  -0.907 0.364777
## endbuyerD                  -5.030e-01  3.230e-01  -1.557 0.119642
## endbuyerE                  -8.880e-01  3.351e-01  -2.650 0.008131 **
## endbuyerU                  -1.107e+00  3.247e-01  -3.410 0.000666 ***
## year                        1.261e-01  7.266e-03  17.350  < 2e-16 ***
## log(Surface):authorstyle1 -4.493e-02  1.074e-01  -0.418 0.675698
## log(Surface):Interm1        1.090e-01  8.163e-02   1.335 0.182025
## log(Surface):discauth1     -9.104e-02  1.122e-01  -0.811 0.417352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.18 on 1476 degrees of freedom
## Multiple R-squared:  0.6276, Adjusted R-squared:  0.6218
## F-statistic: 108.1 on 23 and 1476 DF,  p-value: < 2.2e-16
```

Looking at the diagnostic plots, our model 1 seems to satisfy the assumptions of linear regression resonablly well. From the residual vs. fitted plot, we see that our residuals are randomly distributed with mean 0. There is no heteroskedacity satisfying the constant variance assumption of linear regression. Our QQ plot appears approximately normal, as well. The residuals vs fitted plot shows that there are no high leverage points, influential points, or outliers. We can see from the summary output that approximately 62.8% of the variation in *logprice* can be explained by our model.

The summary of the final model shows a R^2 of 0.627, which means that 62.7% of the variation in *logprice* can be explained by our model.

4. Summary and Conclusions:

```
## Warning in process_lm(ret, x, conf.int = conf.int, conf.level =
## conf.level, : Exponentiating coefficients, but model did not use a log or
## logit link function.
```

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

First of all, all the interactions are not that important but still have influence on the price. The most important variables are *Surface*, *lrgfont*, *prevcoll*, *origin_cat*, *finished* ,*dealer*, *year*. The median price is 132.1425733.

Then, holding all other variables constant,

- We expect 10% increase in *Surface* will increase the price by 3.4% ($1.1^{\hat{\beta}_1} - 1$). We are 95% confident the increase is between about 2.8% and 4.0%. Specifically, if the authors name is introduced, a 10% increase in *Surface* will result in 3% increase of the price; If there is an intermediary involved, a 10%

Table 2: Coefficient Summary for Final Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.000	12.903	-17.130	0.000	0.000	0.000
log(Surface)	1.423	0.029	12.294	0.000	1.345	1.506
lrgfont1	2.306	0.121	6.919	0.000	1.819	2.922
Interm1	1.081	0.472	0.166	0.868	0.428	2.731
authorstyle1	0.454	0.631	-1.250	0.212	0.132	1.568
prevcoll1	2.261	0.143	5.720	0.000	1.709	2.991
origin_catF	0.530	0.076	-8.318	0.000	0.456	0.615
origin_catI	0.483	0.106	-6.838	0.000	0.392	0.595
origin_catO	0.429	0.110	-7.665	0.000	0.346	0.533
engraved1	2.042	0.144	4.964	0.000	1.540	2.708
finished1	2.228	0.091	8.817	0.000	1.864	2.663
discauth1	2.867	0.660	1.595	0.111	0.785	10.472
dealerL	3.662	0.129	10.083	0.000	2.845	4.714
dealerP	1.371	0.159	1.988	0.047	1.004	1.871
dealerR	5.998	0.103	17.357	0.000	4.899	7.344
lands_sc1	0.655	0.116	-3.645	0.000	0.522	0.822
endbuyerC	0.744	0.325	-0.907	0.365	0.393	1.410
endbuyerD	0.605	0.323	-1.557	0.120	0.321	1.140
endbuyerE	0.411	0.335	-2.650	0.008	0.213	0.794
endbuyerU	0.330	0.325	-3.410	0.001	0.175	0.625
year	1.134	0.007	17.350	0.000	1.118	1.151
log(Surface):authorstyle1	0.956	0.107	-0.418	0.676	0.774	1.180
log(Surface):Interm1	1.115	0.082	1.335	0.182	0.950	1.309
log(Surface):discauth1	0.913	0.112	-0.811	0.417	0.733	1.138

increase will result in 4.47% increase of the price; If the dealer engages with the authenticity of the painting, a 10% increase will result in 2.5% increase of the price.

- If the dealer devotes an additional paragraph, the price is expected to increase by 130.6% ($e^{\hat{\beta}_2}$). We are 95% confident the increase is between about 81.8% and 192.2%.
- If an intermediary is involved in the transaction, the price is expected to increase by 8.1%. We are 95% confident the increase is between about -57.2% and 173.1%.
- If the authors name is introduced, the price is expected to decrease by 54.6%. We are 95% confident the decrease is between about -56.8% and 86.8%.
- If the previous owner is mentioned, the price is expected to increase by 126.1%. We are 95% confident the increase is between about -70.9% and 199.1%.
- Compared to Dutch/Flemish, French origin is expected to lead to 47% decrease in price, and we are 95% confident the decrease is between about 38.5% and 54.4%; Italian is expected to lead to 51.7% decrease in price, and we are 95% confident the decrease is between about 40.5% and 60.8%; Other and Spanish origin is expected to lead to 57.1% decrease in price, and we are 95% confident the decrease is between about 46.7% and 55.4%.
- If the dealer mentions engravings done after the painting, the price is expected to increase by 104.2%. We are 95% confident the increase is between about 54.0% and 170.8%.
- If the the painting is finished, the price is expected to increase by 122.8%. We are 95% confident the increase is between about 86.4% and 166.3%.
- If the dealer engages with the authenticity of the painting, the price is expected to increase by 186.7%. We are 95% confident the increase is between about -21.5% and 947.2%.

- Compared to dealer J, dealer L is expected to lead to 266.2% increase in price, and we are 95% confident the decrease is between about 184.5% and 371.5%; Dealer P is expected to lead to 37.1% increase in price, and we are 95% confident the decrease is between about 0.4% and 87.1%; Dealer R is expected to lead to 499.8% increase in price, and we are 95% confident the decrease is between about 389.8% and 634.4%.
- If the painting is described as a plain landscape, the price is expected to decrease by 34.5%. We are 95% confident the decrease is between about -17.8% and 47.8%.
- Compared to a buyer endbuyer, a collecter endbuyer is expected to lead to 25.6% decrease in price, and we are 95% confident the decrease is between about -41.0% and 60.7%; A dealer endbuyer is expected to lead to 39.5% decrease in price, and we are 95% confident the decrease is between about -14.0% and 67.9%; An expert endbuyer is expected to lead to 58.9% decrease in price, and we are 95% confident the decrease is between about 20.6 and 78.3%; An unknown endbuyer is expected to lead to 77.0% decrease in price, and we are 95% confident the decrease is between about 37.5% and 82.5%.
- On average, one unit increase in year will lead to 13.4% increase in price. e are 95% confident the decrease is between about 11.8% and 15.1%.

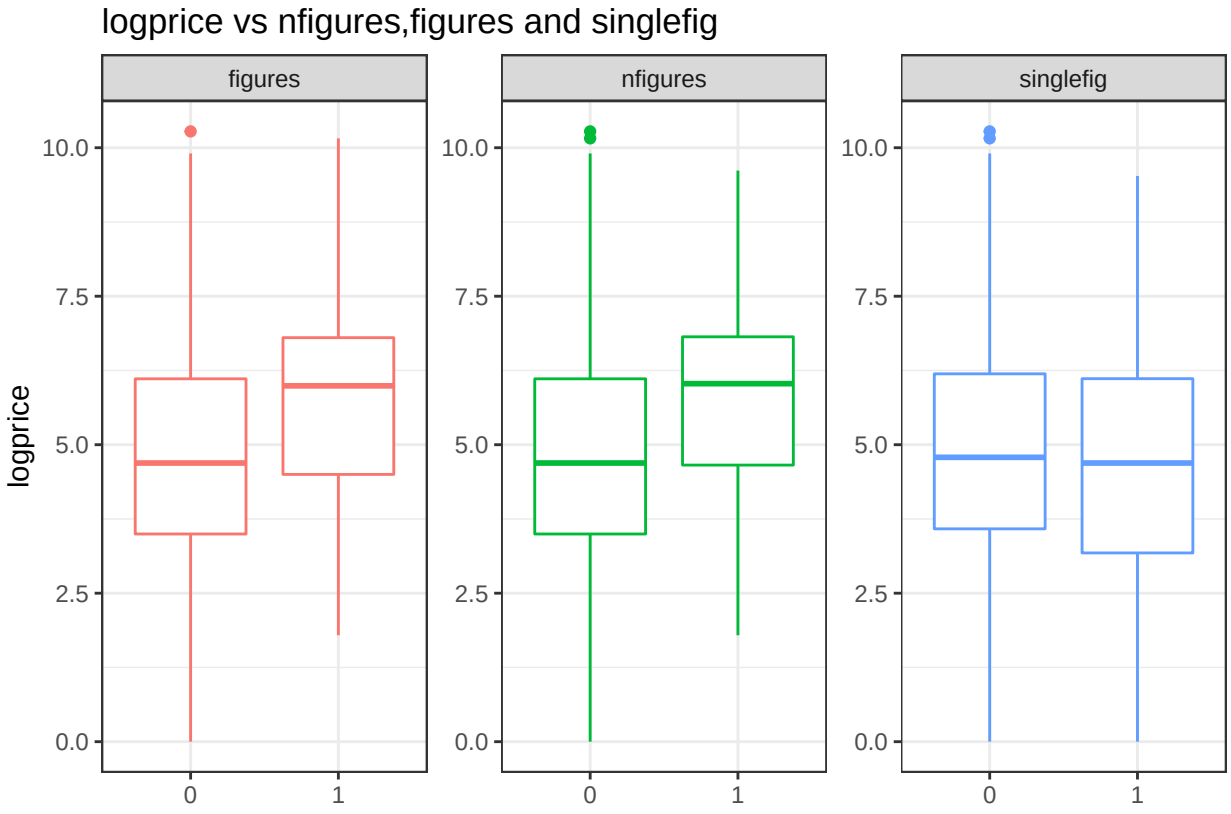
Based on the model, we can make several suggestions to the art gallary:

In our model, the confidence intervals for Surface, Irgfont1, prevcoll1, engraved1, finished1, dealerL, dealerP, dealerR, year strictly exclude 0 and thus, contribute positively to the response variable. However, notice that when surface interact with authorstyle, interm and discauth, confidence intervals for interaction terms are not all postive. Thus, we further combine these terms when making recommendations. One thing to notice is that, considering the fact that terms that are used to interact with Surface are not strictly positve(either not all values in confidence interval are positive or the coefficient itself is negative), we can still take surface as a preferrable feature. This means larger surface size, more recent paintings produced in later years(larger number of year variable), dealer being able to devote an additional paragraph or mentioning engravings done after the painting, previous owner being mentioned, painting being finished, and the dealer of the painting is L,P or R can all contribute positively to the auction price. On the other hand, observing the confidence interval table, we also observe some variables with strictly negative coefficents. Due the fact that most of our variables are categorical, we might consider the base level of those variables have positve effect on auction price. For instance, all levels listed for origin_cat have confidence intervals that only includes negative values, which means having the origin of painting based on dealers' classification in the catalogue being Dutch/Flemish can help increase auction price. In addition, if the painting is not described as a plain landscape, it can be more valuable in terms of auction price for a similar reason. Thus, paintings with such combined features tend to have higher auction price according to our model.

Limitations

There are a few limitations of our model. A major limitation is not having sufficient coverage for test predictions. We only achieved a 65% coverage rate for our 95% prediction interval. This might be imprived when we use more complex model and include more varaibles in our model, like *Author*. Another limitation is the year variable. While this linear regression might work for this short time frame, this analysis suggests that older paintings are worth less than newer paintings. However, this doesn't capture the 'antique market' for classic pieces of art, which are extremely expensive. This limitation is clear when observing the intercept. The intercept assumes the year is 0, which is not logical for this analysis.

Appendix


logprice vs nfigures,figures and singlefig
figures
nfigures
singlefig
logprice
10.0
7.5
5.0
2.5
0.0
0
1