Part-I-Writeup

1. Introduction:

In this study, we are looking at the auction prices of paintings in 18th century Paris. Specifically, through the assistance of model built based on existing training data, we wish to understand the factors that drive the prices of the paintings, and then be able to predict auction prices based on characteristics of a certain painting. After fitting appropriate model, we also intend to detect specific paintings that are either underpriced or overpriced based on the selected model.

One of the main task and challenge is to narrow down the number of potential predictors from 59 to less than 20 while maintaining a high performance of the model. But being able to explain the results and provide some recommendations to indivisuals without statistical background is equally important and challenging. Therefore, we aim at balancing the performance of model prediction, closeness to true model, simplicity, and interprebility.

2. Exploratory data analysis:

A) Data summary & cleaning

To start with, we looked at the summary of the original trainig 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 need to be taken care of. The followings steps are how we cleaned the data:

a. The first step we did was to get rid of intuitivelly useless variables, including: `lot`, `sale`, `price`, `count`, `subject`, `authorstandard`, `author`, `winningbidder`, and ‘other. The are not useful in predicting the response variable (such as names)

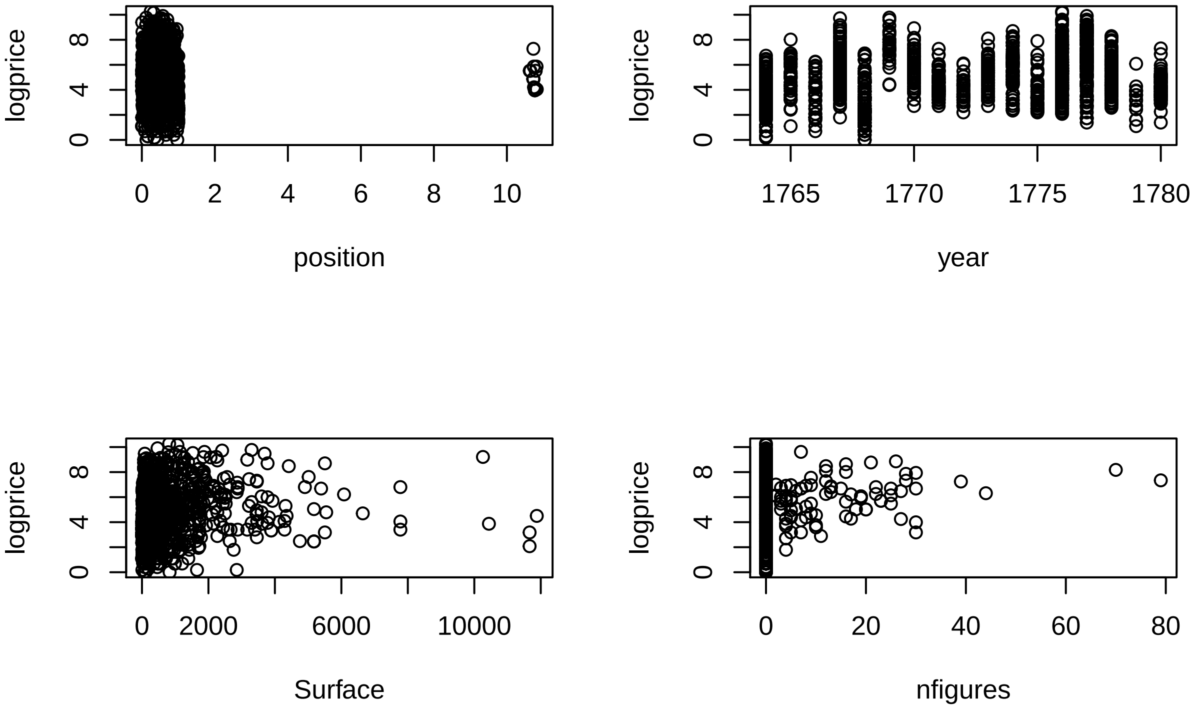
b. By further screening the variables, we found out that `Surface` and `Surface_Rnd`, `Surface_Rect` are corerlated, which are based on the value of `Height_in`, `Width_in`, and `Diam_in`. We decided to use `Surface` in our initial model. The same issue happened to `material`, `mat`, and `materialCat`. The latter one recodes the previous one. Therefore, we used `materialCat`. We applied the same strategy to keep `landsALL` and get rid of other variables related with landscape.

c. 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 “no information”. For `materialCat` and `Shape`, since there are so many levels, we grouped some levels with few observations together, coded as “other” group. The rest binary vairables are changed into factor.

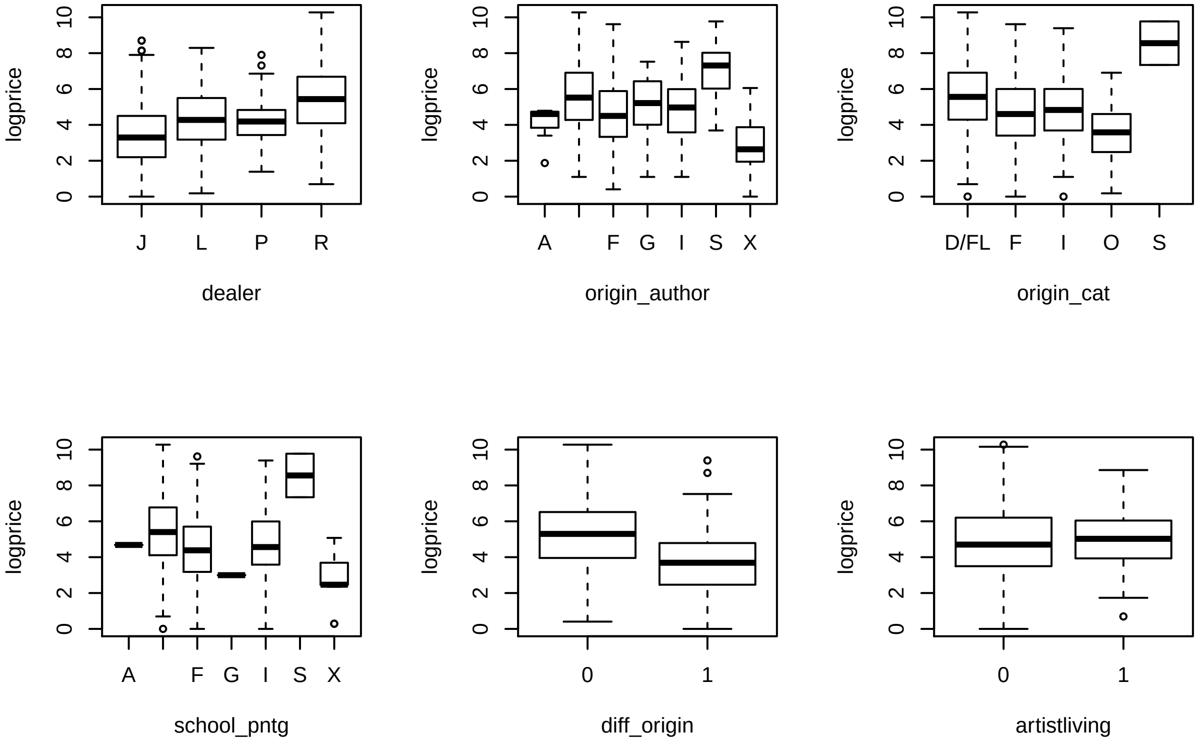
d. Then we dealt with the missing values in `Surface` and `Interm`. We used the package “mice” to address this problem, which uses the observed values in the dataset to impute the missing values. It prevents directly throwing away the missing values, which results in lossing a large amont of information for prediction.

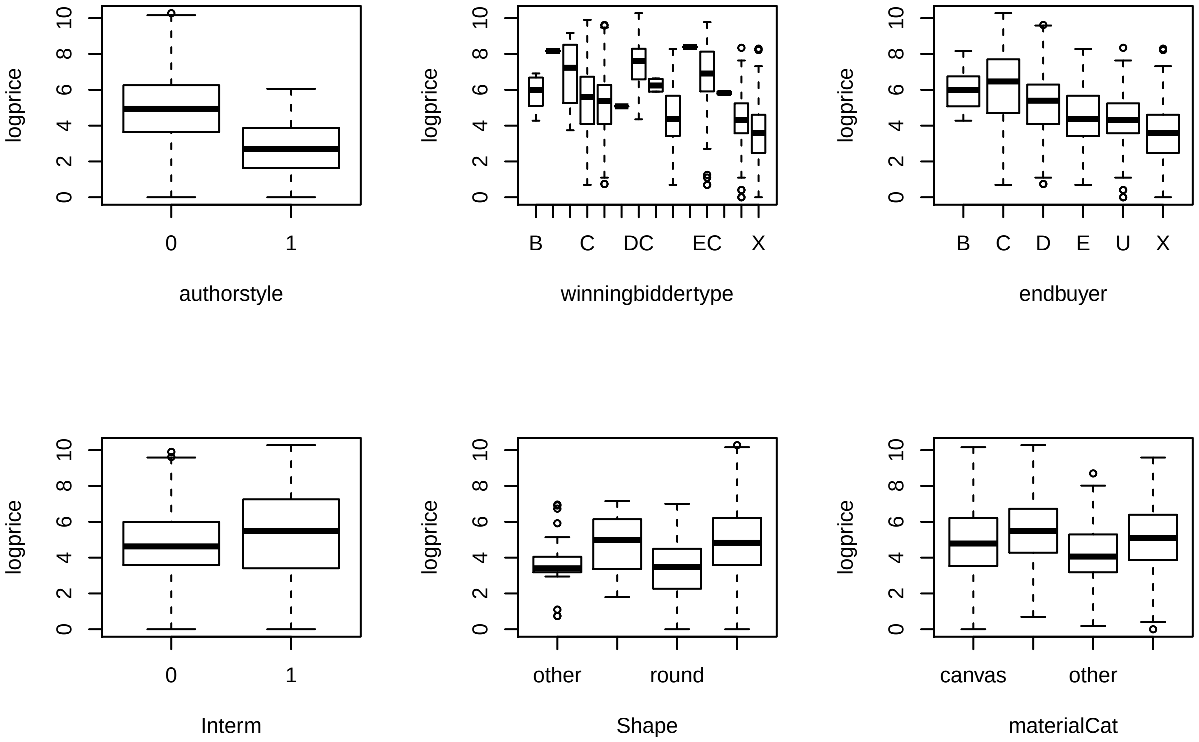
B). Plots

Then we analyed the relationship between those left features and the response variable. With the scatter plots, we can roughly determine which variables can be put into the initial model. For categorical variables, we want to check if the `logprice` spans different ranges in different levels. 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 some 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`. But we'll keep it in the model first.





Since there are 33 categorical variables, we don't show the boxplots for all of them. But applied the same method to check all the categorical variables. The following variables show some differences in `logprice` at different levels (not considering the magnitude of the difference at this time): `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 10 best predictive variables for predicting, we would consider the magnitude of differences and the strength of relationships. The 10 variables we choose are: `Surface`, `dealer`, `school_pntg`, `diff_origin`, `authorstyle`, `endbuyer`, `Interm`, `prevcoll`, `engraved`, `Irgfont`.

3. Development and assessment of an initial model

Initial Model

JZS prior

```
##  [1] "Intercept"           "dealerL"             "dealerR"
##  [4] "year"                "school_pntgD/FL"     "school_pntgS"
##  [7] "school_pntgX"        "diff_origin1"        "artistliving1"
## [10] "authorstyle1"        "winningbiddertypeC"  "winningbiddertypeDD"
## [13] "winningbiddertypeX"  "endbuyerD"           "endbuyerE"
## [16] "endbuyerU"           "Interm1"             "Shaperound"
## [19] "Surface"             "materialCatother"    "materialCatwood"
## [22] "engraved1"           "prevcoll1"           "paired1"
## [25] "finished1"           "lrgfont1"            "portrait1"
## [28] "still_life1"         "discauth1"
```

g-prior

```
##  [1] "Intercept"            "dealerL"              "dealerR"
##  [4] "year"                 "school_pntgD/FL"      "diff_origin1"
##  [7] "artistliving1"        "authorstyle1"         "winningbiddertypeD"
## [10] "winningbiddertypeDC" "winningbiddertypeDD" "winningbiddertypeU"
## [13] "endbuyerE"            "endbuyerX"            "Interm1"
## [16] "Shapeoval"            "Shaperound"           "Surface"
## [19] "materialCatother"     "engraved1"            "prevcoll1"
## [22] "paired1"              "finished1"            "lrgfont1"
## [25] "portrait1"            "still_life1"          "discauth1"
```

The EDA process gives us an initial idea of which variables to drop out to reduce the dimension, and which variables might be significant in explaining the variation in logprice. But before we built the initial model, we applied BMA, Bayesian Model Averaging, to systemetically choose which base variables that have higher posterior probabilities to be in the initial model. We experimented two modelpriors, “JZS” and “g-prior”, which gave us two sets of variables listed above. Then we picked up the common ones from Best Predictive Model(BPM).

Then we fit the linear regression model using the chosen features and all their possible interactions. From the summary table, the $R^2 = 0.5828$, which is fairly high. But we realized that lots of estimated coefficients for interactions are NAs, indicating that some levels in those variables have too few observations to be estimated. Therefore, we need to further reduce the dimention through variable selection.

```
##
## Call:
## lm(formula = logprice ~ dealer + school_pntg + diff_origin +
##     artistliving + endbuyer + authorstyle + Interm + Shape +
##     Surface + engraved + prevcoll + paired + finished + lrgfont +
##     portrait + discauth + still_life, data = paintings_train_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7280 -0.7709  0.0583  0.7968  4.6960
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.471e+00  1.350e+00   3.313 0.000947 ***
## dealerL          1.487e+00  1.364e-01  10.902  < 2e-16 ***
## dealerP          5.226e-01  1.667e-01   3.135 0.001752 **
## dealerR          1.172e+00  1.090e-01  10.757  < 2e-16 ***
## school_pntgD/FL -5.929e-01  1.271e+00  -0.467 0.640903
## school_pntgF    -1.336e+00  1.272e+00  -1.051 0.293495
## school_pntgG    -2.952e+00  1.791e+00  -1.649 0.099455 .
## school_pntgI    -1.284e+00  1.274e+00  -1.008 0.313772
## school_pntgS     3.029e-02  1.562e+00   0.019 0.984528
## school_pntgX    -2.060e+00  1.285e+00  -1.603 0.109058
## diff_origin1    -6.790e-01  9.240e-02  -7.349 3.29e-13 ***
## artistliving1    6.357e-01  1.057e-01   6.011 2.32e-09 ***
## endbuyerC       -5.417e-02  3.467e-01  -0.156 0.875846
## endbuyerD       -1.815e-01  3.441e-01  -0.528 0.597881
## endbuyerE       -8.624e-01  3.582e-01  -2.408 0.016180 *
## endbuyerU       -6.788e-01  3.539e-01  -1.918 0.055317 .
## endbuyerX       -1.573e+00  3.523e-01  -4.464 8.67e-06 ***
## authorstyle1    -1.074e+00  1.588e-01  -6.761 1.97e-11 ***
## Interm1          5.793e-01  9.683e-02   5.982 2.76e-09 ***
## Shapeoval        6.532e-01  3.908e-01   1.672 0.094832 .
## Shaperound      -9.857e-02  3.561e-01  -0.277 0.781987
## Shapesqu_rect    8.292e-01  2.601e-01   3.188 0.001463 **
## Surface          1.516e-04  3.206e-05   4.731 2.45e-06 ***
## engraved1        3.464e-01  1.528e-01   2.267 0.023562 *
## prevcoll1        1.235e+00  1.516e-01   8.150 7.72e-16 ***
## paired1         -3.248e-01  7.085e-02  -4.584 4.94e-06 ***
## finished1        6.088e-01  9.756e-02   6.240 5.70e-10 ***
## lrgfont1         1.199e+00  1.239e-01   9.678  < 2e-16 ***
## portrait1       -6.097e-01  1.779e-01  -3.427 0.000627 ***
## discauth1        3.945e-01  1.461e-01   2.700 0.007019 **
## still_life1     -7.400e-01  1.730e-01  -4.278 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 1469 degrees of freedom
## Multiple R-squared:  0.5775, Adjusted R-squared:  0.5689
## F-statistic: 66.93 on 30 and 1469 DF,  p-value: < 2.2e-16
```

Model Selection

After completing the initial exploratory data analysis, methods including Stepwise Best Subset Selection using both AIC and BIC were used in order to assess more systematically which covariates were most important for predicting the logprice of paintings. While the number of relevant covariates was initially thinned by examining the data and determining which variables were best suited for modeling (e.g. via dimension reduction, elimination or recoding of categorical variables with too many levels or too few observations for a given level to be useful in estimating a coefficient), there still remained a large number of covariates from which to choose. The goal in using the above described methodology was to demonstrate among several methods, both frequentist and Bayesian, which covariates were routinely deemed to be the most important for modeling logprice.

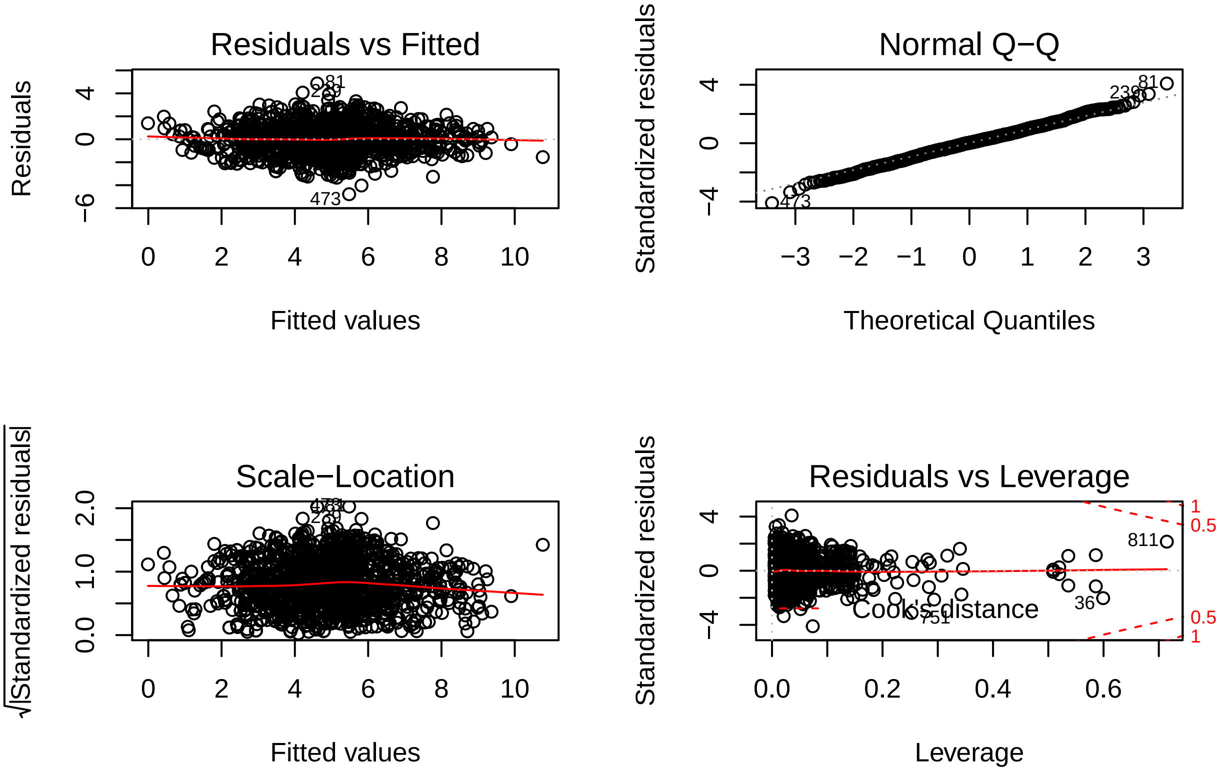
The variable selection methods described above remain computationally intensive, particularly given the number of variables and potential two-way interactions that must be considered. In order to begin the analysis, The two-way interactions were considered using stepwise selection (AIC & BIC). The goal of this penalized selection process was to avoid overfitting and to deliver a model that was both interpretable and performed well in prediction. Then we compared the results from two methods and filtered out interactions that have NAs as coefficients, that are not significant, and that do not make sense to be interacted (such as $artistling * endbuyer$).

Ultimately, the following variables were selected using the above methods and were fit using OLS regression. The R^2 reduces to 0.6315, which is expected. All the estimated coefficients do not contain NAs.

```
##
## Call:
## lm(formula = logprice ~ Shape + school_pntg + dealer * Interm +
##     dealer * Surface + dealer * paired + dealer * finished +
##     dealer * discauth + diff_origin * Surface + diff_origin *
##     portrait + artistliving * endbuyer + artistliving * authorstyle +
##     Interm * Surface + Interm * lrgfont + Surface * lrgfont +
##     Surface * still_life + Surface * discauth + prevcoll * finished +
##     paired * lrgfont + paired * discauth + diff_origin * authorstyle +
##     diff_origin * still_life + finished * discauth + lrgfont *
##     discauth + artistliving * finished + Interm * portrait, data = paintings_train_2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.7874 -0.7301  0.0265  0.7492  4.8638 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                4.080e+00  1.338e+00   3.049 0.002335 ** 
## Shapeoval                  8.227e-01  3.844e-01   2.140 0.032501 *  
## Shaperound                -2.032e-01  3.520e-01  -0.577 0.563854    
## Shapesqu_rect              7.779e-01  2.610e-01   2.980 0.002928 ** 
## school_pntgD/FL           -6.346e-01  1.240e+00  -0.512 0.608985    
## school_pntgF              -1.308e+00  1.242e+00  -1.053 0.292495    
## school_pntgG              -3.059e+00  1.745e+00  -1.753 0.079795 .  
## school_pntgI              -1.297e+00  1.244e+00  -1.043 0.297321    
## school_pntgS              -5.197e-01  1.539e+00  -0.338 0.735684    
## school_pntgX              -1.905e+00  1.256e+00  -1.517 0.129511    
## dealerL                    2.514e+00  2.206e-01  11.394  < 2e-16 ***
## dealerP                    1.161e+00  2.839e-01   4.088 4.59e-05 ***
## dealerR                    1.717e+00  1.867e-01   9.193  < 2e-16 ***
## Interm1                    5.570e-02  2.589e-01   0.215 0.829703    
## Surface                    7.240e-04  1.927e-04   3.757 0.000179 ***
## paired1                    1.716e-01  1.910e-01   0.899 0.369014    
## finished1                  1.067e+00  2.248e-01   4.746 2.28e-06 ***
## discauth1                  9.838e-01  3.963e-01   2.482 0.013167 *  
## diff_origin1              -5.498e-01  1.068e-01  -5.149 2.98e-07 ***
## portrait1                 -8.675e-01  2.090e-01  -4.150 3.52e-05 ***
## artistliving1             -5.993e-01  7.969e-01  -0.752 0.452192    
## endbuyerC                 -3.338e-01  3.802e-01  -0.878 0.380169    
## endbuyerD                 -3.956e-01  3.764e-01  -1.051 0.293398    
## endbuyerE                 -1.182e+00  3.895e-01  -3.035 0.002449 ** 
## endbuyerU                 -9.123e-01  3.871e-01  -2.357 0.018575 *  
## endbuyerX                 -1.787e+00  3.847e-01  -4.645 3.71e-06 ***
## authorstyle1              -1.818e+00  4.294e-01  -4.233 2.45e-05 ***
## lrgfont1                   1.654e+00  1.776e-01   9.315  < 2e-16 ***
## still_life1               -2.032e-01  2.446e-01  -0.831 0.406192    
## prevcoll1                  1.493e+00  1.715e-01   8.702  < 2e-16 ***
## dealerL:Interm1            2.589e-01  3.009e-01   0.860 0.389721    
## dealerP:Interm1            1.130e-01  6.790e-01   0.166 0.867883    
## dealerR:Interm1            8.495e-01  2.808e-01   3.025 0.002528 ** 
## dealerL:Surface           -5.943e-04  2.206e-04  -2.694 0.007143 ** 
## dealerP:Surface           -2.222e-04  3.453e-04  -0.643 0.520127    
## dealerR:Surface           -4.802e-04  1.943e-04  -2.471 0.013583 *  
## dealerL:paired1           -1.205e+00  2.553e-01  -4.721 2.57e-06 ***
## dealerP:paired1           -6.697e-01  3.560e-01  -1.881 0.060168 .  
## dealerR:paired1           -2.608e-01  2.097e-01  -1.243 0.213974    
## dealerL:finished1         -5.267e-01  4.988e-01  -1.056 0.291151    
## dealerP:finished1         -6.354e-01  4.136e-01  -1.536 0.124652    
## dealerR:finished1         -5.409e-01  2.459e-01  -2.200 0.027987 *  
## dealerL:discauth1          1.084e-02  9.377e-01   0.012 0.990774    
## dealerP:discauth1         -1.787e+00  9.636e-01  -1.854 0.063896 .  
## dealerR:discauth1         -9.134e-01  3.690e-01  -2.475 0.013428 *  
## Surface:diff_origin1      -1.391e-04  8.487e-05  -1.639 0.101489    
## diff_origin1:portrait1     9.767e-01  4.184e-01   2.334 0.019711 *  
## artistliving1:endbuyerC    8.808e-01  8.233e-01   1.070 0.284903    
## artistliving1:endbuyerD      1.096e+00  8.134e-01   1.348 0.177943
## artistliving1:endbuyerE      2.074e+00  8.705e-01   2.382 0.017333 *
## artistliving1:endbuyerU      1.222e+00  8.375e-01   1.459 0.144688
## artistliving1:endbuyerX      1.395e+00  8.205e-01   1.700 0.089278 .
## artistliving1:authorstyle1  1.043e+00  8.965e-01   1.163 0.244990
## Interm1:Surface              5.230e-05  1.044e-04   0.501 0.616432
## Interm1:lrgfont1            -2.324e-01  2.527e-01  -0.920 0.357959
## Surface:lrgfont1            -4.772e-05  1.071e-04  -0.446 0.655968
## Surface:still_life1         -3.139e-04  2.394e-04  -1.311 0.189990
## Surface:discauth1           -1.026e-04  2.197e-04  -0.467 0.640702
## finished1:prevcoll1         -9.668e-01  3.300e-01  -2.930 0.003448 **
## paired1:lrgfont1            -8.417e-01  2.582e-01  -3.260 0.001142 **
## paired1:discauth1           -3.384e-01  3.503e-01  -0.966 0.334210
## diff_origin1:authorstyle1   9.401e-01  4.595e-01   2.046 0.040937 *
## diff_origin1:still_life1    -9.031e-01  3.548e-01  -2.546 0.011009 *
## finished1:discauth1          5.219e-01  3.153e-01   1.655 0.098123 .
## discauth1:lrgfont1          -4.018e-01  4.426e-01  -0.908 0.364190
## finished1:artistliving1     -2.925e-01  2.856e-01  -1.024 0.305978
## Interm1:portrait1           -4.275e-01  4.296e-01  -0.995 0.319780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.213 on 1433 degrees of freedom
## Multiple R-squared:  0.6179, Adjusted R-squared:  0.6004
## F-statistic: 35.12 on 66 and 1433 DF,  p-value: < 2.2e-16
```

Residuals



After fitting the model, we created the four model diagnostic plots. The overall appearances of all four plots seem acceptable, with no obvious outlier or highly influential points shown. The model also does not violate the normality assumption. The constant variance of residuals assumption seems to be satisfied. However, there are 2 cases that are dropped from the plots because they both have leverage of 1, indicating that they could potentially be the outlying cases of underpriced/overpriced paintings that we will later on investigate in, or have extreme price values. It is worth our attention to specifically look at these cases.

Variables

	Coefficient	2.5%	97.5%
(Intercept)	59.168	4.287	816.686
Shapeoval	2.277	1.071	4.839
Shaperound	0.816	0.409	1.628
Shapesqu_rect	2.177	1.305	3.633
school_pntgD/FL	0.530	0.047	6.041
school_pntgF	0.270	0.024	3.090
school_pntgG	0.047	0.002	1.439
school_pntgI	0.273	0.024	3.137
school_pntgS	0.595	0.029	12.179
school_pntgX	0.149	0.013	1.748
dealerL	12.350	8.012	19.039
dealerP	3.192	1.829	5.571
dealerR	5.567	3.859	8.030
Interm1	1.057	0.636	1.757
Surface	1.001	1.000	1.001
paired1	1.187	0.816	1.727
finished1	2.906	1.870	4.517
discauth1	2.675	1.229	5.819
diff_origin1	0.577	0.468	0.712
portrait1	0.420	0.279	0.633
artistliving1	0.549	0.115	2.622
endbuyerC	0.716	0.340	1.510
endbuyerD	0.673	0.322	1.409
endbuyerE	0.307	0.143	0.658
endbuyerU	0.402	0.188	0.858
endbuyerX	0.167	0.079	0.356
authorstyle1	0.162	0.070	0.377
lrgfont1	5.230	3.691	7.409
still_life1	0.816	0.505	1.319
prevcoll1	4.449	3.178	6.228
dealerL:Interm1	1.295	0.718	2.338
dealerP:Interm1	1.120	0.296	4.241
dealerR:Interm1	2.338	1.348	4.056
dealerL:Surface	0.999	0.999	1.000
dealerP:Surface	1.000	0.999	1.000
dealerR:Surface	1.000	0.999	1.000
dealerL:paired1	0.300	0.182	0.494
dealerP:paired1	0.512	0.255	1.029
dealerR:paired1	0.770	0.511	1.163
dealerL:finished1	0.591	0.222	1.571
dealerP:finished1	0.530	0.235	1.192
dealerR:finished1	0.582	0.359	0.943

	Coefficient	2.5%	97.5%
dealerL:discauth1	1.011	0.161	6.361
dealerP:discauth1	0.167	0.025	1.109
dealerR:discauth1	0.401	0.195	0.827
Surface:diff_origin1	1.000	1.000	1.000
diff_origin1:portrait1	2.656	1.169	6.034
artistliving1:endbuyerC	2.413	0.480	12.132
artistliving1:endbuyerD	2.993	0.607	14.758
artistliving1:endbuyerE	7.955	1.442	43.874
artistliving1:endbuyerU	3.395	0.657	17.549
artistliving1:endbuyerX	4.035	0.807	20.176
artistliving1:authorstyle1	2.837	0.489	16.466
Interm1:Surface	1.000	1.000	1.000
Interm1:lrgfont1	0.793	0.483	1.301
Surface:lrgfont1	1.000	1.000	1.000
Surface:still_life1	1.000	0.999	1.000
Surface:discauth1	1.000	0.999	1.000
finished1:prevcoll1	0.380	0.199	0.727
paired1:lrgfont1	0.431	0.260	0.715
paired1:discauth1	0.713	0.359	1.417
diff_origin1:authorstyle1	2.560	1.040	6.306
diff_origin1:still_life1	0.405	0.202	0.813
finished1:discauth1	1.685	0.908	3.128
discauth1:lrgfont1	0.669	0.281	1.594
finished1:artistliving1	0.746	0.426	1.307
Interm1:portrait1	0.652	0.281	1.515

In the linear model we selected, we included `Shape`, `school_pntg`, `dealer`, `Interm`, `Surface`, `paired`, `finished`, `discauth`, `diff_origin`, `portrait`, `artistliving`, `endbuyer`, `authorstyle`, `lrgfont`, `still_life`, and `prevcoll` as our base predictors. Interactions selected by the model selection process and, for the sake of interpretation, those that are reasonable and interpretable are kept in the model as well. Since the response variable was orginally log-transformed, the exponentiated coefficients and confidence intervals are shown in the table.

4. Summary and Conclusions

```
##        fit      lwr      upr
## 1 72.22846 5.256927 992.3955
```

```
##        fit      lwr      upr
## 1 72.22846 2.097641 2487.056
```

What is the (median) price for the "baseline" category if there are categorical or dummy variables in the model (add CI's)? (be sure to include units!) Highlight important findings and potential limitations of your model. Does it appear that interactions are important? What are the most important variables and/or interactions? Provide interprations of how the most important variables influence the (median) price giving a range (CI). Correct interpretation of coefficients for the log model desirable for full points.

Provide recommendations for the art historian about features or combination of features to look for to find the most valuable paintings.

a. The median price predicted is `exp(4.401232) = 81.55128` livres. The 95% confidence interval is (6.248, 1064.357) livres. The prediction interval is (2.532, 2626.714) livres.

Interpretation

From the final model we fitted, we found out that the following variables are statistically significant: `dealer`, `Interm`, `Surface`, `finished`, `discauth`, `diff_origin`, `portrait`, `endbuyer (E,U, X)`, `authorstyle`, `lrgfont`, and `prevcoll`. Some of the interactions are statistically important, such as: `dealer*Interm`, `dealer*paried`, `Interm*lrgfont`, `diff_origin*portrait` etc. We picked the most important ones and interpreted as following:

- dealerL: compared with dealer J, the average selling price from dealer L is `exp(2.526) = 12.50339` times higher. (Same interpretation for dealer P and R, with different coefficients)
- Interm: when there is an intermediary involved in the transaction, the selling price is `exp(-1.523) = 0.218` times lower than when there is no intermediary involved.
- Surface: for every one squared inches increase in the painting surface, the selling price is expected to increase 8.779e-4 livres.
- finished: if the painting is finished, the selling price on average is `exp(1.087) = 2.965` times higher than when the painting is not finished.
- portrait: if the painting is described as a portrait, the selling price on average is `exp(-0.9246) = 0.3967` times lower than when the painting is not described as a portrait.
- dealerR&Interm1: if the dealer is R and with an intermediary existed, the average selling price is 1.854 times higher when the dealer is J with an intermediary.

Recommendations

In order to find the most valuable paintings, we recommend historians focusing on several features (just to mention some, not a complete list). For example, they might want to look for dealer R, with an intermediary involved; they might want to look for dealer L with the painting sold as a pair with another one; they should look for larger, finished paintings; they should focus on paintings whose authors' names are mentioned during the auction.