Part-II-Writeup

Team-FP03

2019/12/15

Introduction

In this project, we are going to explore what factors drove the price of paintings in 18th century Paris, and thus to identify possible overvalued and undervalued paintings.

The dataset we are going to analyze is a series of auction transactions of paintings in Paris, ranging from 1764 to 1780. This dataset mainly contains the following information:

1. Sale data, this include basic information about dealers, end buyers, transaction dates and prices;

2. Characteristics of paintings, such as their painters, sizes, materials, number of figures and themes.

To address our problem, we divide this project into three parts:

1. In the first part, we carried out an exploratory data analysis. The target of this section is to understand the composition of our dataset and identify potential important variables.

2. In the second part, a simple linear regression model was fit to the data, aiming to confirm important variables and interactions from the model selection process and to prepare for fitting a more complex model.

3. In the last part, we attempt a range of different models on the dataset, challenging ourselves to achieve a better prediction model. At last, we decide to use a random forest model as our final choice mainly due to its satisfactory accuracy.

Exploratory Data Analysis

In this section, we are going to explore our dataset in the following way: we first investigate the variables in the dataset to find their characteristics and possible relationships among each other; then we check the scatter plots between the response and each variable to identify potential important predictors.

Variable investigation

The dataset has 1500 observations and 59 variables in total, among which 19 are of type character and 40 are recognized as type numeric by R. However, with the following analysis, it's easy to find that a considerable amount of them should be considered as categorical varialbes, which is also mentioned in the provided codebook.

Variables to drop

First of all, we can remove a few variables from the list of potential predictors simply based on their definitions:

Variable `price` is just the exponetial form of our target response `logprice`, and thus needs removing;
Variable `count` is the same for all observations, therefore there's no point to use it in the model fitting.

Besides these two, there exist quite a number of variables of interest:

Variables `subject`, `lot`, and `material` have way too many distinct values. Also, the possible values for these variables are too complicated and we decide not to use them in the model.

Then, in the dataset there exist strong correlations among some pairs of variables. For example, there is correlation between `Interm` & `type_intermed`, and `mat` & `materialCat`. In **Table 1**, we display the contingency table for `Interm` vs. `type_intermed`, and as we can see, when `Interm` takes 0 `type_intermed` always takes `n/a`; when `Interm` takes 1, `type_intermed` takes other values. Thus, we decide to remove `Interm`, since `type_intermed` hopefully contains a bit more information. This is similar for the relationship between `materialCat` and `mat`, and thus we remove `materialCat`.

Table 1: Interm vs type_intermed

		B	D	E	EB
0	960	0	0	0	0
1	0	11	94	39	1

In a similar manner, `Surface` should be known if `Diam_in`, or `Height_in` and `Width_in` are known at the same time. Also, note that `Surface` is the combination of `Surface_Rnd` and `Surface_Rect`. Thus, among all these variables mentioned, we just keep `Surface` in the model fitting process.

Addtionally, if variables `origin_author` and `origin_cat` are known, the value of `diff_origin` is also 100% certain. Besides, `type_intermed` incorporates all information of `Interm`, and `sale` is just the combination of `dealer` and `year`, but has way more distinct values which is not easy to handle. Thus, we decide to drop `diff_origin`, `Interm` and `sale`.

Variables to impute

We've found that NA's exist in a lot of variables, and these NA's do not always indicate values missing completely at random. For example, from the R output below, we can see that `Surface` is not missing at random, since the coefficient for the categorical variable `is.na(Surface)` is judged to be sigificant with a p-value close to 0.

Thus, for these numeric variables, instead of simply discarding observations containing NA's, we choose to impute the missing values with the observed ones.

For categorical variables with a lot of blank or meaningless values, such as `endbuyer`, `type_intermed`, `material` and `mat`, we impute `n/a` into them to create a new category.

```
##
## Call:
## lm(formula = paintings_train$logprice ~ is.na(paintings_train$Surface))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9691 -1.3316 -0.0978  1.2455  5.5980
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        4.96915    0.04969 100.002   <2e-16
## is.na(paintings_train$Surface)TRUE -1.86766    0.21383  -8.734   <2e-16
##
## (Intercept)                        ***
## is.na(paintings_train$Surface)TRUE ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.872 on 1498 degrees of freedom
## Multiple R-squared:  0.04846,    Adjusted R-squared:  0.04782
```

```
## F-statistic: 76.29 on 1 and 1498 DF,  p-value: < 2.2e-16
```

Variables to manipulate

Variable `position` indicates the position of lot in the catalogue and is expressed as percentages. However, the maximum value of it in the dataset can be as large as 10.82, which are obviously typos. Similarly, there are observations with variables such as `Surface` equal to 0. As a result, observations with impossible `position` and `Surface` values are dropped.

Besides, `Shape` variable has some weird values, such as `oval` vs. `ovale`, and `ronde` vs. `round`, which are also typos and thus need fixing. Similarly, for variable `authorstyle` there are values like `in the taste`, `taste of` and `in the taste of` which essentially mean the same thing. For variables like this, we incorporate these redundant levels.

Variables `authorstandard` and `author` seem to have prohibitively many distinct values. However, it is almost commmon sense that the value of a painting should be correlated with its author, especially with famous authors. Thus, we managed to create a new variable `Fame`, which indicates whether the author for a specific painting has created a masterpiece that ranked high in terms of logarithmic sale price.

In order to determine the range of famous painters, we take reference to the following plot **Figure 1**, which displays the trend of `logprice` across the dataset. It seems that there's a very sharp decrease of `logprice` for expensice paintings at the beginning of the plot, which is very likely to be caused by 'celebrity effect'. Thus, we decide to choose the authors of top 120 paintings to be famous, since after roughly 120 the decrease becomes much smoother.

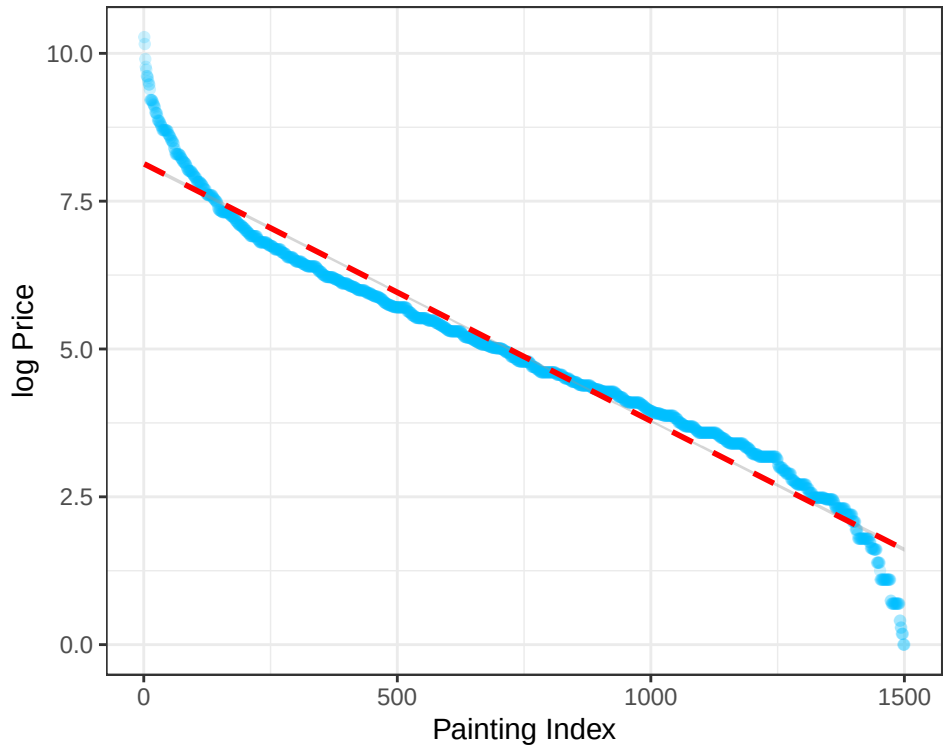

Figure 1: Plot of logprice

At last, notice that `winningbiddertype` has too many levels, which may result in difficulties both in model fitting and in interpretation. Thus, we decide to apply the following transformation on `winningbiddertype`:

Observations with levels `B,BB,BC` are combined to have level `B`;
Observations with levels `C` remains untouched;
Observations with levels `D,DB,DC,DD` are combined to have level `D`;
Observations with levels `E,EB,EBC,EC,ED` are combined to have level `E`;
Blank space and unknown observations are combined to have level `n/a`.

The rationale for the above transformation is that, the bidder who actually attended the auction had the most influence on the sale price.

Important predictor identification

In this section we are going to evaluate scatter plots between our response `logprice` and each varaible after the manipulations from the previous part.

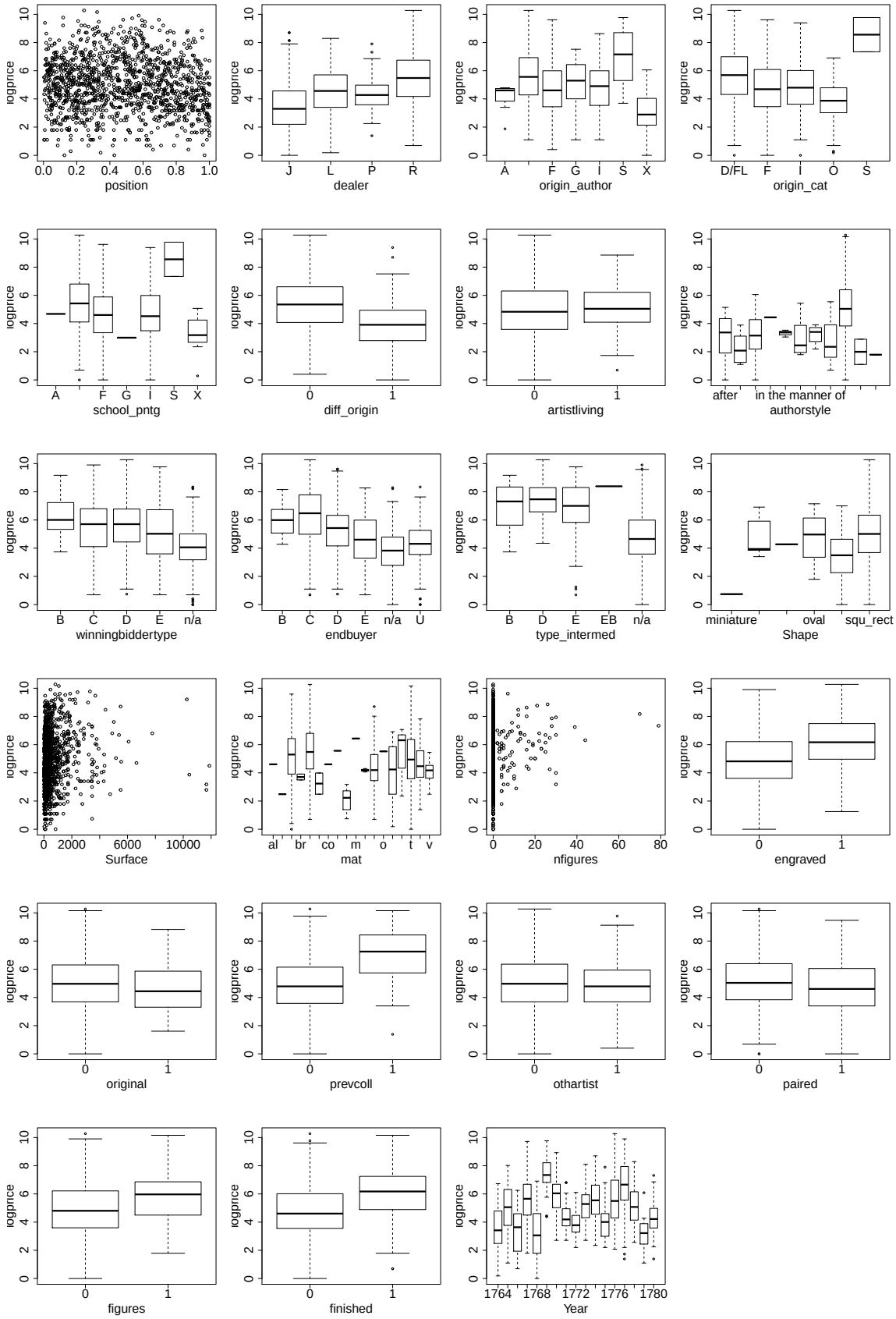

Figure 2: Plots of predictors versus logprice (1 to 23)

Figure 2 above displays the scatter plots between `logprice` and the first 23 variables in the dataset. Our target is to identify variables that show a strong relationship with the response. Bearing this in mind, it is easy to notice that variables `dealer`, `origin_author`, `winningbiddertype`, `endbuyer`, `type_intermed`, `prevcoll`, `finished` and `year`appear to have the strongest relationship with `logprice`. In addition, variables such as `Surface` are clustered near the beginning of x axis, and thus we decide to apply log transformations on them and have a closer look afterwards.

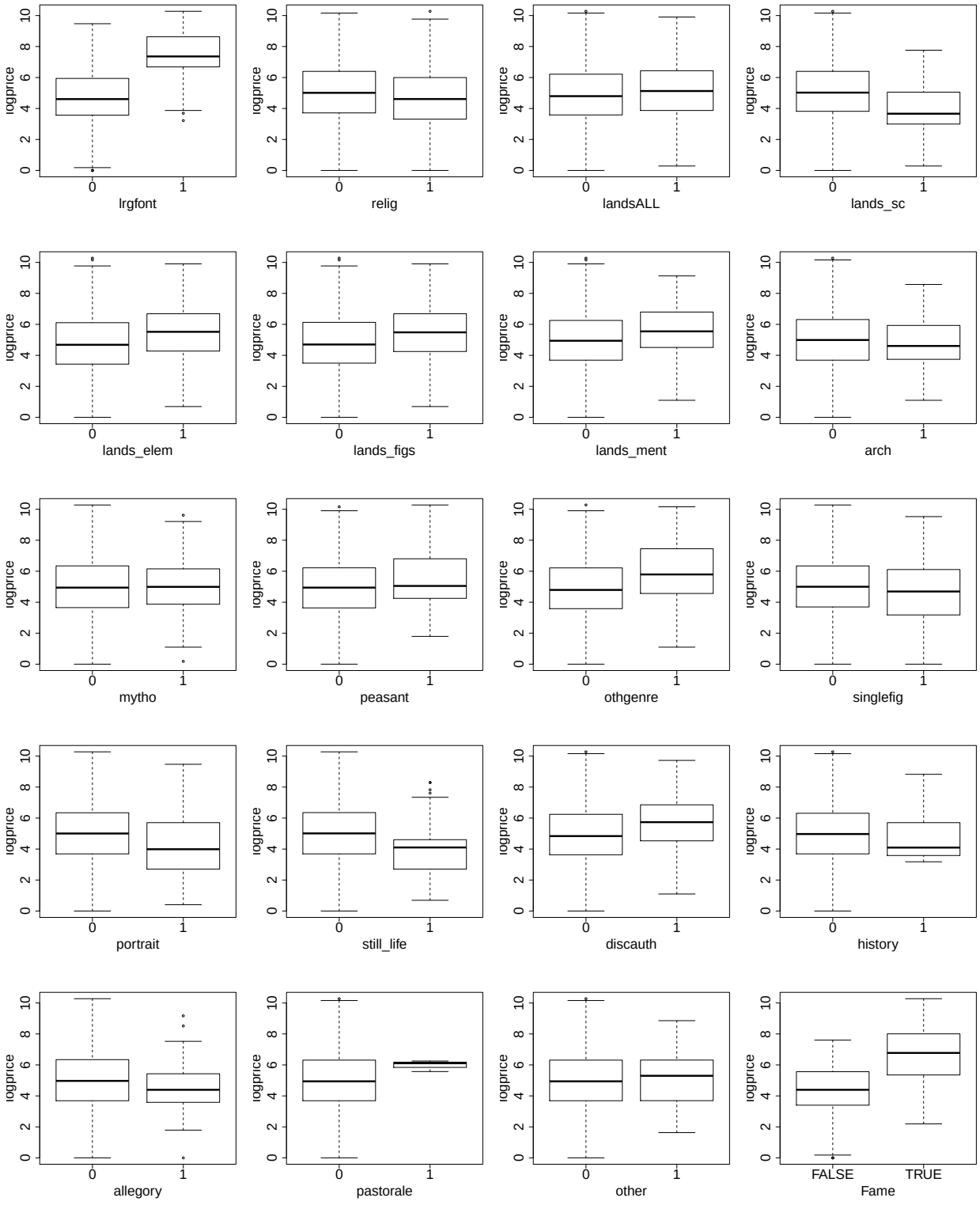

Figure 3: Plots of predictors versus logprice (24 to 42)

Figure 3 above display the scatter plots between `logprice` and the rest of the variables in the dataset. As we can see, most of the binary categorical variables fail to present a strong relationship with the response. The only exceptions are `lrgfont` and the created variable `Fame`, which correspond to quite different response values at the two different levels.

For `Surface`, we can do log transformation to the corresponding predictors to see their relationship with `logprice` at a greater detail in **Figure 4**.

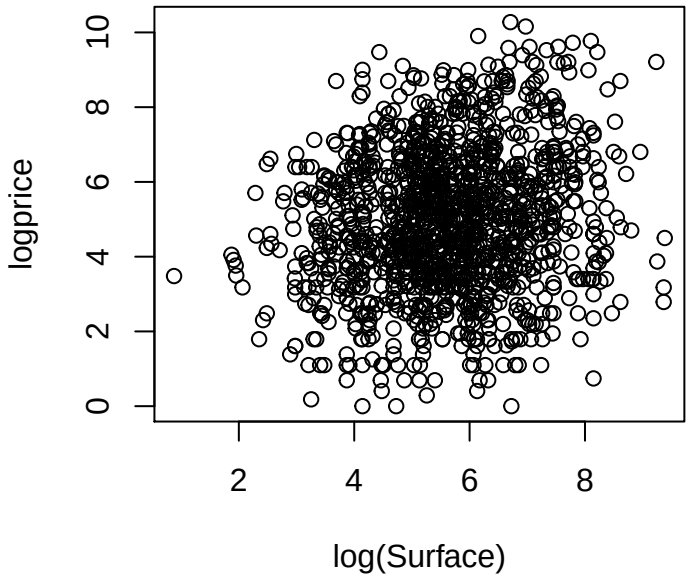

Figure 4: Plots of log Surface versus logprice

As we can see from **Figure 4**, there seem to be a weak relationship between `logprice` and log-transformed `Surface`. Intuitively, the surface of paintings should indeed be correlated to their prices.

In conclusion, after our manipulation of the dataset and inspection of the relationships between response and each variable, we reckon that variables `dealer`, `year`, `origin_author`, `winningbiddertype`, `endbuyer`, `type_intermed`, `finished`, `lrgfont`, `Fame` and the log transformation of `Surface` are the most important variables in terms of scatter plots and their definitions. To be specific, these variables represent the following meaning:

1. `dealer` indicates dealer initials;
2. `year` means the year of sale;
3. `origin_author` indicates the origin of painting based on nationality of artist;
4. `winningbiddertype` is the type of winning bidder, such as buyer on behalf of collector or expert;
5. `endbuyer` is the type of end buyer;

6. `type_intermed` means the type of intermediary;

7. `finished` indicates whether the painting is noted for its highly polished finishing;

8. `lrgfont` indicates whether the dealer devotes an additional paragraph (always written in a larger font size);

9. `Surface` is the surface of painting in squared inches, and is transformed on log scale;

10. `Fame`, which is a created variable, indicates whether a painter has created one of the top 120 expensive paintings in the dataset.

In addition, variables like `prevcoll` and `school_pntg` also seem important, and may be of use as well. However, we need formal model fitting and selection process to decide the variables and interactions to use.

Preliminary Model Discussion

In this section, we are going to discuss the performance of our preliminary model fitted in part I.

Discussion of preliminary model

The results of our linear regression model on the leaderboard based on the original test data are shown in the table below.

As we can see, our `Bias` result is relatively high, while `Coverage` and `RMSE` results are relatively low. This is not a very satisfactory model. Moreover, owing to the property of linear regression, our potential to improve the performance is very limited. Despite the fact that `Bias` is likely to decrease if more predictors are added to the model, `Coverage` and `RMSE` are most likely to increase in the meantime, leading to no significant improvement of our model. This relationship is also known as the `Bias-Variance` tradeoff. As a result, we can hardly improve our result under linear regression model. Therefore, it is necessary to introduce some non-linear regression models, such as Random Forest.

Table 2: Preliminary Model Performance

Type	Bias	Coverage	MaxDeviation	MeanAbsDeviation	RMSE
linear	206	0.95	13558	541	1487

Final Model fitting

In this section, we are going to present our final model, which is a random forest model built upon the experience from EDA and the preliminary model fitting process and after attempting a series of different models.
The remaining part of the section will proceed in the following way: we will first display the summary table for our final model, and then provide expalnations for the variables employed in the model. After these, we will explain in detail the model selection process, including what standards are used in variable selection and the rationales behind. Then a residual plot resulted from our final model will be displayed and appropriate descriptions will be offered. At last, we will provide a clear procedure explaining how the prediction interval was obtained, since there are no explicit methods for acquiring a prediction interval for random forest models.

Development of the final model

The summary table of our final model is displayed below:

Table 3: Summary Table - Variables

variable
year
log_Surface
dealer
lrgfont
endbuyer
origin_author
winningbiddertype
finished
type_intermed
diff_origin
precvoll
paired
Fame
mat

Table 4: Summary Table - Model

No. of trees	mtry	Variance Explained	Mean of squared residuals
500	13	0.672	1.178

As is seen from the summary table, we have used the following variables:

1. `year`: year of sale;
2. `log_surface`: the logarithmic transformation of surface of paintings in squared inches;
3. `dealer`: a categorical variable representing dealer initials with 4 unique dealers: J, L, P and R;
4. `lrgfont`: indicates whether the dealer devotes an additional paragraph (always written in a larger font size);
5. `endbuyer`: a categorical variable indicating the type of end buyer, with B = buyer, C = collector, D = dealer, E = expert organizing the sale, X = identity unknown and blank = no information;
6. `origin_author`: origin of painting based on nationality of artist, with A = Austrian, D/FL = Dutch/Flemish, F = French, G = German, I = Italian, S = Spanish and X = Unknown;
7. `winningbiddertype`: indicating the type of winning bidder wih B = buyer, C = collector, D = dealer, E = experts organization and n/a = no information;
8. `finished`: indicating whether the painting is finished, with '1' indicating painting is finished;
9. `type_intermed`: a categorical variable representing the type of intermediary with B = buyer, D = dealer and E = expert;
10. `diff_origin`: indicating whether variable `origin_author` is different from `origin_cat`; in other words, it means whether the origin of the paintings based on nationality and dealer's classification are the same or not, with 1 representing the same;

11. `Fame`: indicating whether the author of the painting is famous, with '1' indicating that the author is famous;

12. `paired`: indicating whether the painting is sold or suggested as a pairing for another, with '1' indicating it's sold as a pairing for another;

13. `mat`: representing the category of material, with 'al' = alabaster, 'ar' = slate, 'b' = wood, 'br' = bronze frames, 'c' = copper, 'ca' = cardboard, 'co' = cloth, 'g' = grissaille technique, 'h' = oil technique, 'm' = marble, 'mi' = miniature technique, 'o' = other, 'p' = paper, 'pa' = pastel, 't' = canvas, 'ta' = canvas, 'v' = glass and n/a = NA;

14. `prevcoll`: indicating if the previous owner of the painting is mentioned, with '1' indicating yes.

During the variable selection process, we selected variables based on the following three criteria.

The first criterion is the linear model in **Preliminary Model Fitting** part. Our fitted linear model chose some of the important variables identified in EDA, and according to test data accuracy and coverage, it performed satisfactorily. Thus, we also consider some of the variables used in the previous part.
The second criterion is the **variable importance plot** resulted from our random Forest model, as displayed above. Besides those variables already selected from the first criterion, we also selected some of the important variables in the top of the importance plot.
The third criterion is the overall performance as indicated by the leadboard. We elaborated our model to achieve a higher performance from the leadboard, emphasizing on the test data root mean square error and model coverage.

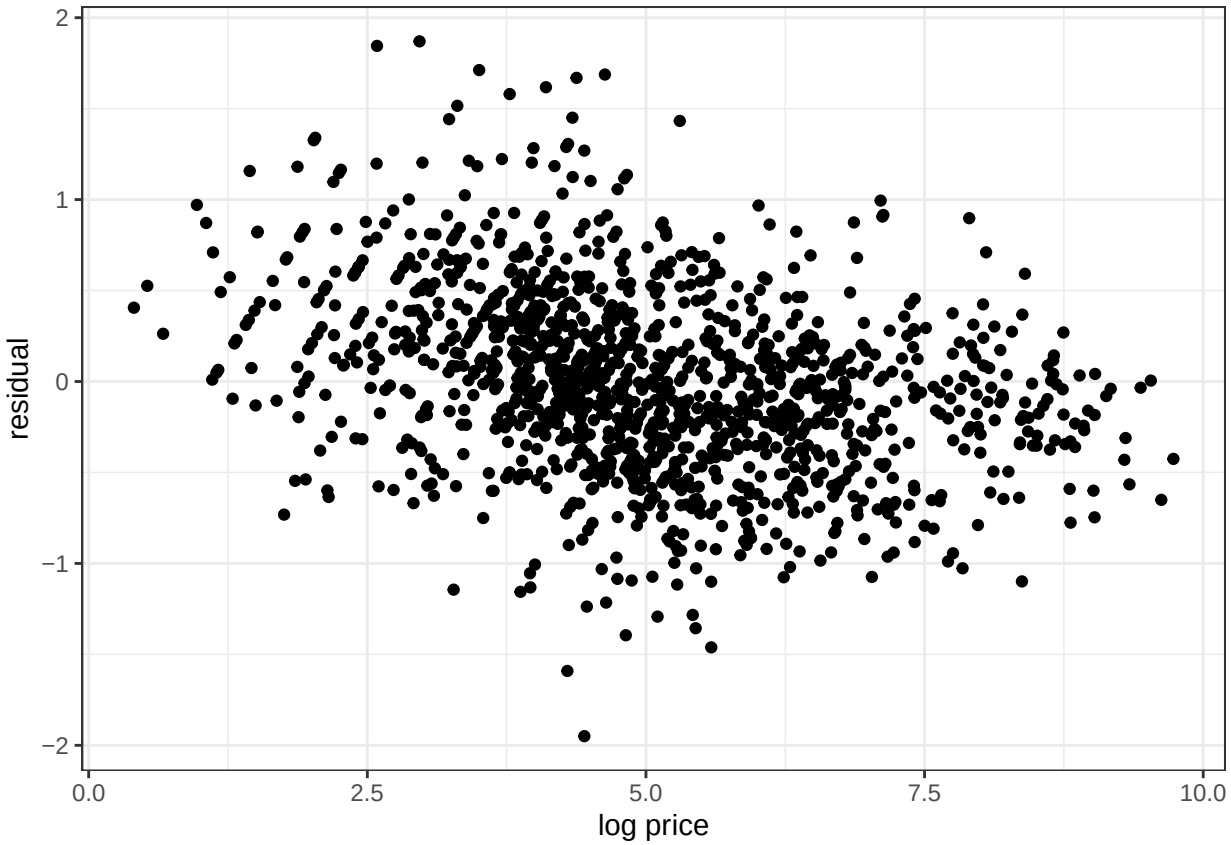

Figure 5: Residual Plot

As we can see from the figure above, the residuals behave better when `logprice` is either very small or very large than the ones when logprice is around 5. More specifically, the residuals have a larger spread (from -2 to 2) when logprice is roughly 5. On the other hand, residuals are relativly tightly centerd around 0 when `logprice` goes into extremes. Generally speaking, most of the residuals are smaller than 2 and are distributed in a reasonable way, which does not indicate any significant problem of our model.

Finally, we are going to explain the method of constructing the prediction interval out of our final random forest model. Since there are no direct funtions nor any explicit methods to calculate prediction intervals for random forest models, we choose to use bootstrap method to calculate the interval.

To build our interval, we first bootstrap our training data to obtain a series of 'new' datasets of the same size. Then, we fit our random forest model on each new dataset and calculate predictions based on test data. Putting all the calculated predictions, we can then acquire the confidence interval. At last, we subtract/add 1.96*rmse (rmse of the training data) to the confidence interval to calculate the prediction interval.

Assessment of the final Model

In this section, we are going to display the following topics: we would first discuss the evaluation of our random forest model based on both the training set and test set; then we are going to display our model result, which includes a selection and discussion of the top 10 valued paintings in the validation data.

Table 5: Performance on training data

bias	coverage	Max AbsDeviation	Mean AbsDeviation	RMSE
-177.348	1	13871.5	248.105	871.8

As we can see, when using our model to predict training data, RMSE is 871.7996258 and coverage is 100%. The coverage is quite satisfying. Also, when we upload our predict-test.Rdata, Wercker shows that for test data, our RMSE is around 886.10, which is similar to training RMSE. This indicates that our model is a good one to reduce overfitting. In addition, covarage for test data is 93.46%, which means that our model can fairly deal with model uncertainty.

Table 6: Top 10 valued paintings

author	year	dealer	surface	endbuyer
Federico Barocci	1768	J	832.00	
Robert Tournières	1768	J	1394.00	
Flemish	1768	J	1386.00	
Johann Wilhelm Baur	1768	J	945.00	U
Flemish	1768	J	30.25	C
Guido Reni	1768	J	180.00	U
Johann Wilhelm Baur	1768	J	154.00	E
Johann Wilhelm Baur	1764	L	192.00	
Anonymous	1768	J	48.00	D
Antoine Dieu	1768	J	48.00	D

From our result of top 10, except one "anonymous", all authors that appear in top 10 are all very famous painters, so adding a new variable `Fame` is reasonable. Also 9 of 10 paintings are sold in 1768 with J-type dealers. This matches with our random forest model, which exactly shows that `Fame`, `year` and `dealer` are top three important variables.

Table 7: Variable Importance Plot

	%IncMSE
Fame	94.331
dealer	83.071
year	80.737
log_Surface	44.684
endbuyer	40.299

Conclusion

The target of this report is to find out the factors that could affect the price of paintings in 18th century Paris, and thus to identify the potentially overvalued and undervalued paintings. We have tried two methods to build the prediction models: linear regression model and Random Forest.

In the linear regression model, we start with choosing the main predictors from EDA (Exploratory data analysis). And we fit the full model including all of the chosen predictors and the interactions between them. Then we implement the BIC procedure and end up with the model including `dealer`, `year`, `origin_author`, `endbuyer`, `log_Surface`, `finished`, `lrgfont`, `winningbiddertype` and `year:winningbiddertype`.

Table 8: Summary of coefficients and confidence intervals of linear model

	Estimate	CI_Low	CI_Up
(Intercept)	-140.858	-438.098	156.381
dealerL	1.397	1.130	1.665
dealerP	0.007	-0.320	0.334
dealerR	1.702	1.480	1.924
year	0.081	-0.087	0.249
origin_authorD/FL	0.406	-0.506	1.318
origin_authorF	-0.119	-1.031	0.792
origin_authorG	-0.070	-1.092	0.952
origin_authorI	-0.299	-1.226	0.629
origin_authorS	-0.113	-1.301	1.075
origin_authorX	-1.005	-1.940	-0.070
endbuyerC	0.317	-0.642	1.275
endbuyerD	-0.828	-1.813	0.157
endbuyerE	-0.810	-1.867	0.247
endbuyern/a	-43.021	-342.333	256.290
endbuyerU	-42.502	-341.809	256.806
log_Surface	0.325	0.272	0.378
finished1	0.963	0.778	1.149
lrgfont1	1.073	0.827	1.319
winningbiddertypeC	-124.699	-428.110	178.713
winningbiddertypeD	-3.773	-303.505	295.958
winningbiddertypeE	-226.302	-530.473	77.869
year:winningbiddertypeC	0.070	-0.101	0.242
year:winningbiddertypeD	0.002	-0.167	0.172
year:winningbiddertypeE	0.128	-0.044	0.300
year:winningbiddertypen/a	0.024	-0.146	0.193

For every one year after the previous year, we expect that the price of the painting will be $e^{0.09}$ times higher, and we are 95% confident that the fluction is between $e^{-0.08}$ to $e^{0.26}$, which is from 0.92 to 1.30.

Given all other conditions unchanged (eg: same dealer, same year, same origin, etc.), we expect the price of the painting will be e^{-110} times higher if the type of winning bidder is a collector. And we are 95% confident that the price fluction will be between e^{-410} and e^{191} times higher.

Given all other conditions unchanged (eg: same dealer, same year,same origin,etc.), we expect the price of painting will be e^{10} times higher if the type of winning bidder is a dealer. And we are 95% confident that the price fluction will be between e^{-286} and e^{307} times higher.

Given all other conditions unchanged (eg: same dealer, same year,same origin,etc.), we expect the price of painting will be e^{-210} times higher if the type of winning bidder is an expert organizing the sale. And we are 95% confident that the price fluction will be between e^{-511} and e^{91} times higher.

In the random forest model, we reselect the main predictors using Variable Importance Measures in Random Forest as well as EDA results and intuition. Then we fit the random forest model using the selected predictors and the result shows as below.

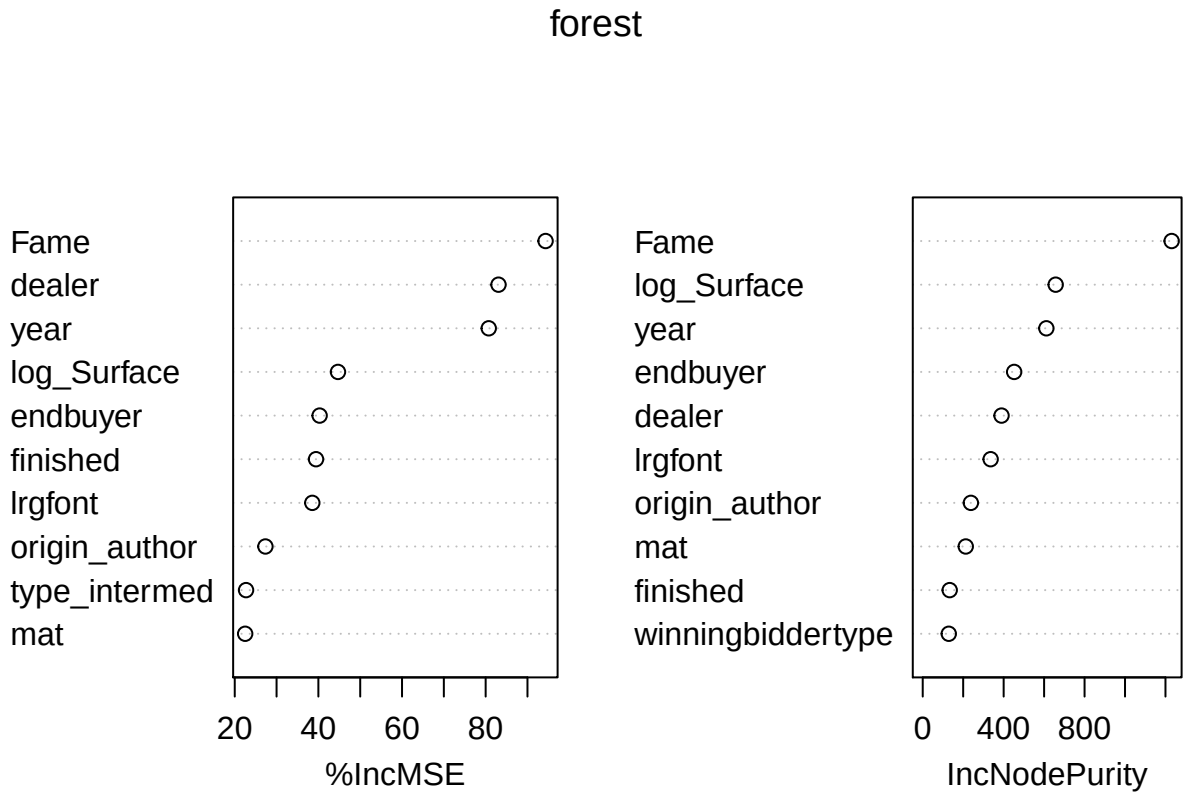

Figure 6: Variable Importance Measures

From the variable importance measures plot of random forest, we can see that from the perspective of accuracy, the 10 most important predictors for random forest are `Fame`, `year`, `dealer`, `log_Surface`, `endbuyer`, `finished`, `lrgfont`, `orgin_author`, `type_intermed` and `winningbiddertype`. These are the factors that could significantly affect the price of paintings in 18th century Paris and are the recommended features to look for to find the most valuable paintings.

The prediction performance of the two models on the new test data is shown below.

Table 9: Summary of Prediction Performance on the Test Data

	linear	rf
Bias	206.171	153.389
Coverage	0.952	0.935
maxDeviation	13558.248	8422.370
MeanAbsDeviation	541.419	313.295
RMSE	1487.499	886.098

Based on the table above, we decide to choose random forest as our final model due to its low `Bias`, similar `coverage`, low `maxDeviation`, low `MeanAbsDeviation` and small `RMSE` compared with linear regression model. But there are also some limitations on our model.

1. Our choice of main predictors is mostly based on random forest, EDA and intuition. It is highly possible

that there exists another combination of predictors that can outperform our current model. If we had more time, we would mainly focus on this problem.

2. Random forest does a good job at classification problems but not as good for regression problems. This is because of its property that it doesn't provide precise continuous prediction. This will lead to the result that random forest may overfit data set when it is particularly noisy.

3. Random forest sometimes feel like a black box and hard to interpret. The users have very little control on what the model does. Sometimes more tries on the parameters and random seeds will provide a better result.