Part-II-Writeup

Team 9

12/14/2019

Introduction

As the second largest city in Europe and the center of art in 18th century, Paris witnessed a lot of auctions on paintings. The most important aspect that all people in auctions would consider is the price. Then the question aries: What could drive the prices of paintings? This project aims to use model fitting methods to find out what factors could affect the price of paintings in 18th century Paris by using the dataset containing information of paintings with auction price data from 1764-1780 on the sales, painters and other characterisitcs of paintings.

In the dataset, each line represents a painting. Each column represents a feature of the painting including the year of sale, width, height, surface area, the author. The original dataset is not clean enough for a statistical analysis. Before we start analyzing, we first cleaned the dataset and impute missing values for several variables by assuming missing completely at random and using "Mice" package that would assign values of missing cells based on existing data. To conduct the analysis, we took the logarithm of prices of each painting and set it as the response variable in all models we tried.

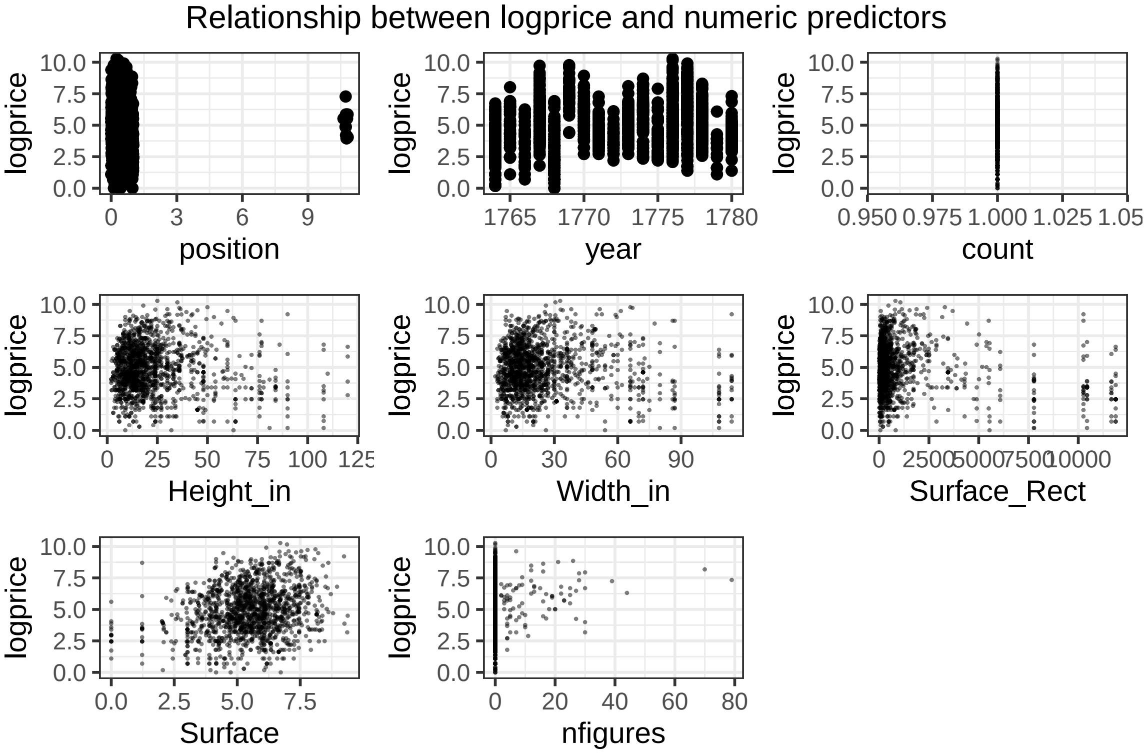
Before implementing different models, we applied exploratory data analysis by drawing scatterplots and boxplots to find out what variables might be related with the price that would give us some ideas of what variables to include in the model. We first fitted a linear model by selecting predictors only based on EDA. We set our initial linear model to contain all selected predictors with all possible two-way interactions. Then we used BIC (Bayesian Information Criterion) to conduct model selection and build a parsimonious model. This simple model would provide us with some ideas about the variables.

Besides only applying results of EDA, we then used Random Forest to choose important variables and then combine results of EDA to form a new set of predictors. Then we apply complex models including simple linear model, BMA (Bayesian Model Averaging), BART (Bayesian Additive Regression Tree), Random Forest, and GAM (Generalized Additive Model) to compare which model could give us the best predictions in terms of RMSE(root mean squared error) and coverage etc.. Predictors and interactions in this final model would give us information about which features or combinations of features would influence the price of paintings so that people could find the most valuable paintings.

Exploratory data analysis

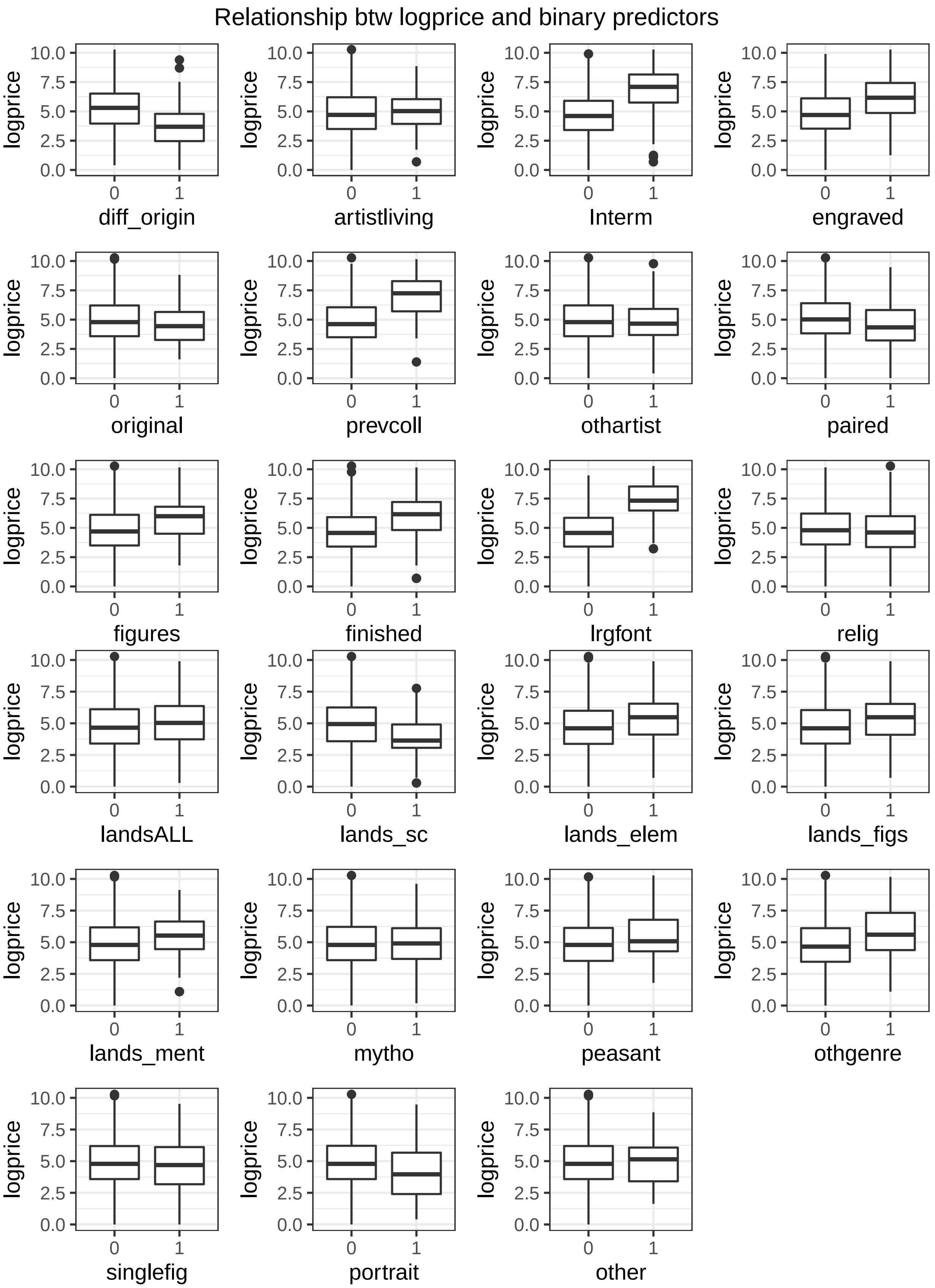
We first checked how many predictors in the original data have missing data. We noticed that `type_intermed`, `Diam_in`, `Surface_Rnd` and `authorstyle` are largely missing, and therefore excluded these predictors from model building. We noticed that many variables have missing or unknown values, and used `mice` to impute the missing values based on the values of the other variables. The following exploratory data analysis is based on the imputed training data set.

To identify the best variables for predicting `logprice`, we want to examine the relationships between `logprice` and the other variables. Since there are too many predictors, we look at quantitative, binary and numeric predictors separately. We first create a scatterplot matrix for the quantitative predictors. We noticed that `position` is in percentage, but some of the paintings have values larger than 1, suggesting that these were recorded wrong. We took log() of `Surface` when cleaning the data, and therefore its distribution is not skewed in the plot. The above plot also shows that, `Height_in`, `Width_in`, `Surface_Rect` need transformation, since their distributions are skewed.



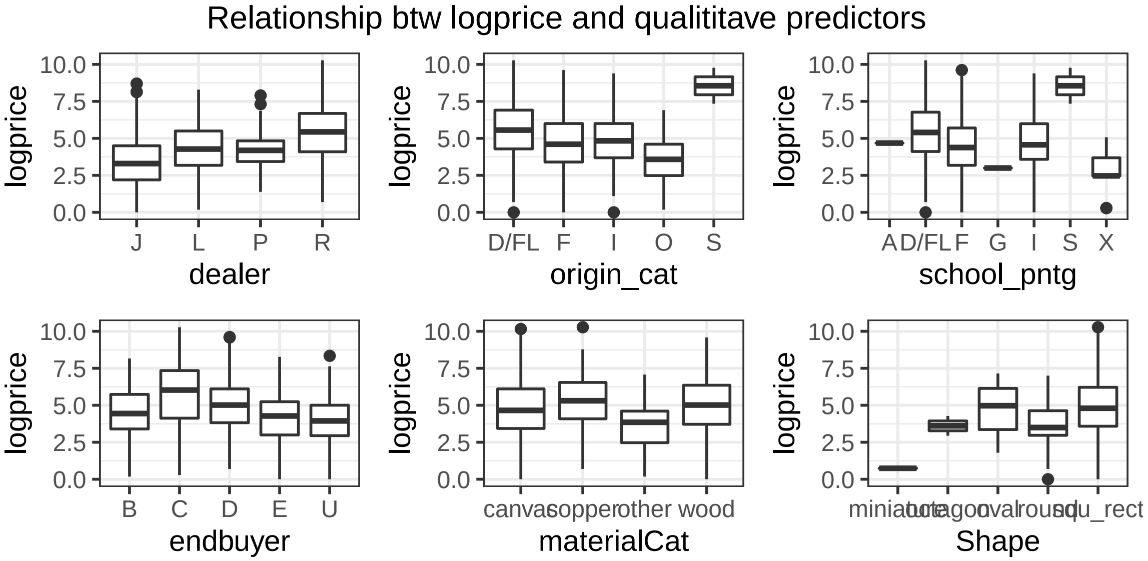
We decided to use `Surface` in our model, because we suspect that paintings with large surface area would be sold at a higher prices. We also wanted to use `year` as a predictor, since the plot shows that `logprice` fluctuates with `year`.

The following is a scatterplot matrix for the binary variables in the data set. The plot shows that we might be able to predict `logprice` based on the variables `diff_origin`, `engraved`, `prevcoll`, `finished`, `lrgfont` and `still_life`.

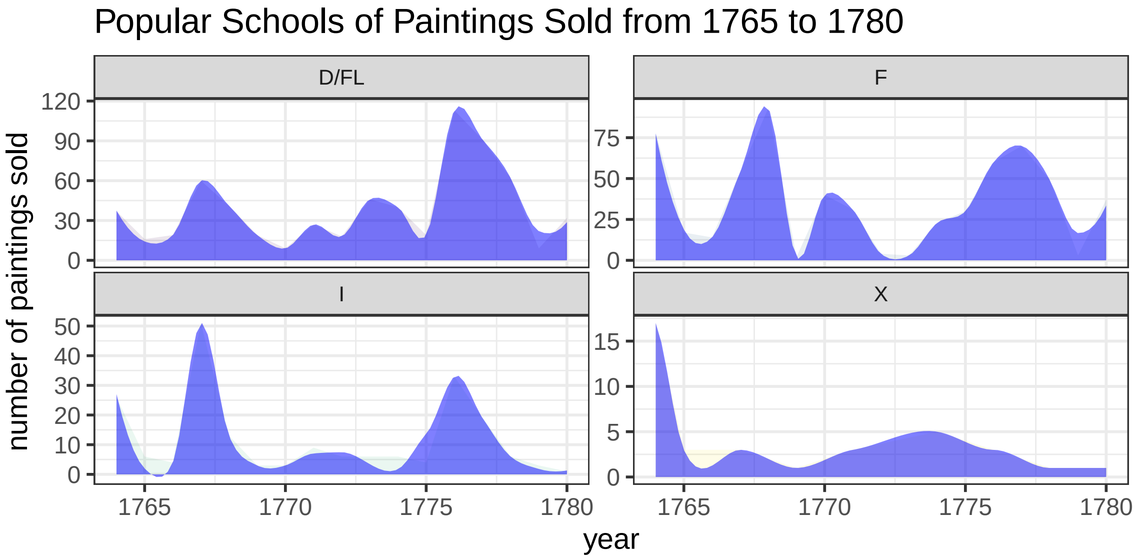


Relationship btw logprice and binary predictors

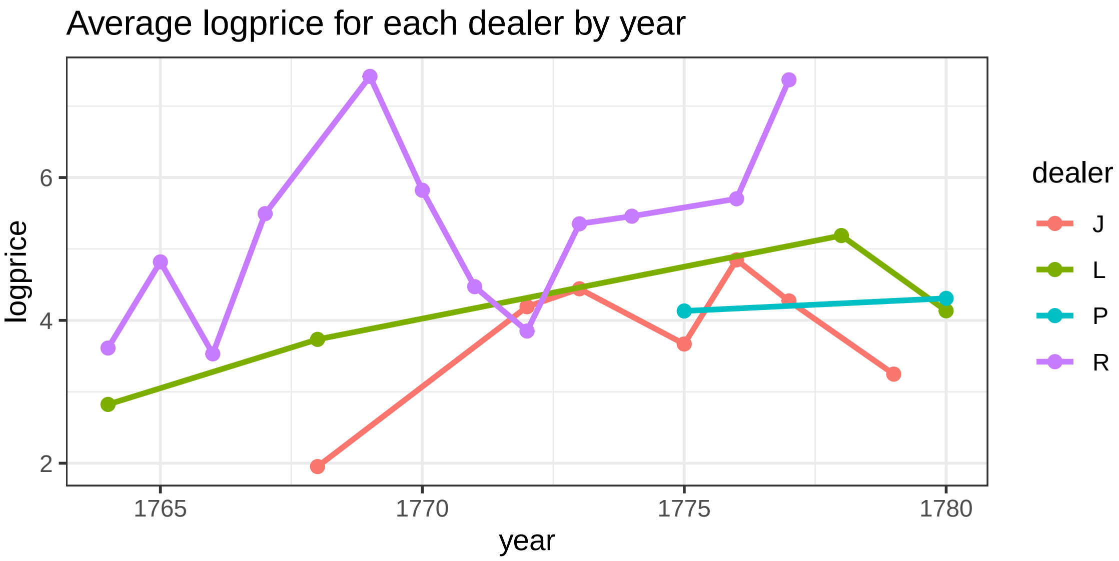
The scatterplot matrix between `logprice` and categorical variables with fewer categories suggest that we could consider all these variables to predict `logprice`.



The following plot investigates the popular school of paintings sold over the years.



The following line graph indicates that dealer R sold paintings at a higher price over the years, while dealer J sold at a lower price over the years.



Beyond what plotted above, there are some variables, such as `lot`, `authorstandard`, `winningbidder`, whose relationships to `logprice` are hard to visualize because they contain too many categories. We decided not to use them from our initial model building. Based on the EDA, we selected the 10 best variables for predicting `logprice`:

- Numerical: `Surface`, `year`
- Binary: `diff_origin`, `engraved`, `prevcoll`, `finished`, `lrgfont`, `still_life`
- Other qualitative: `dealer`, `origin_author`

Discussion of preliminary model Part I

Bias	Coverage	maxDeviation	MeanAbsDeviation	RMSE
263.996	0.953	12960.579	481.893	1269.146

The above table shows the leader board results we got for our OLS model in Part I. Based on the result, our model in Part I already performed really well. Also, according at the summary table, all variables are statistically significantly.

Training.RMSE	Test.RMSE
1513.434	1269.146

By looking at the RMSE for both the training set and the test set, we can further confirm its well performance. Since the model in Part I performed well, we will keep OLS as an option. Further development could be done by choosing variables using Random Forests and fitting into other complex models such as Random Forests, GAM, BAM and BART. We will compare different model results according to their coverage, bias, RMSE, maximum deviation and mean absoulte deviation in order to determine our final model.

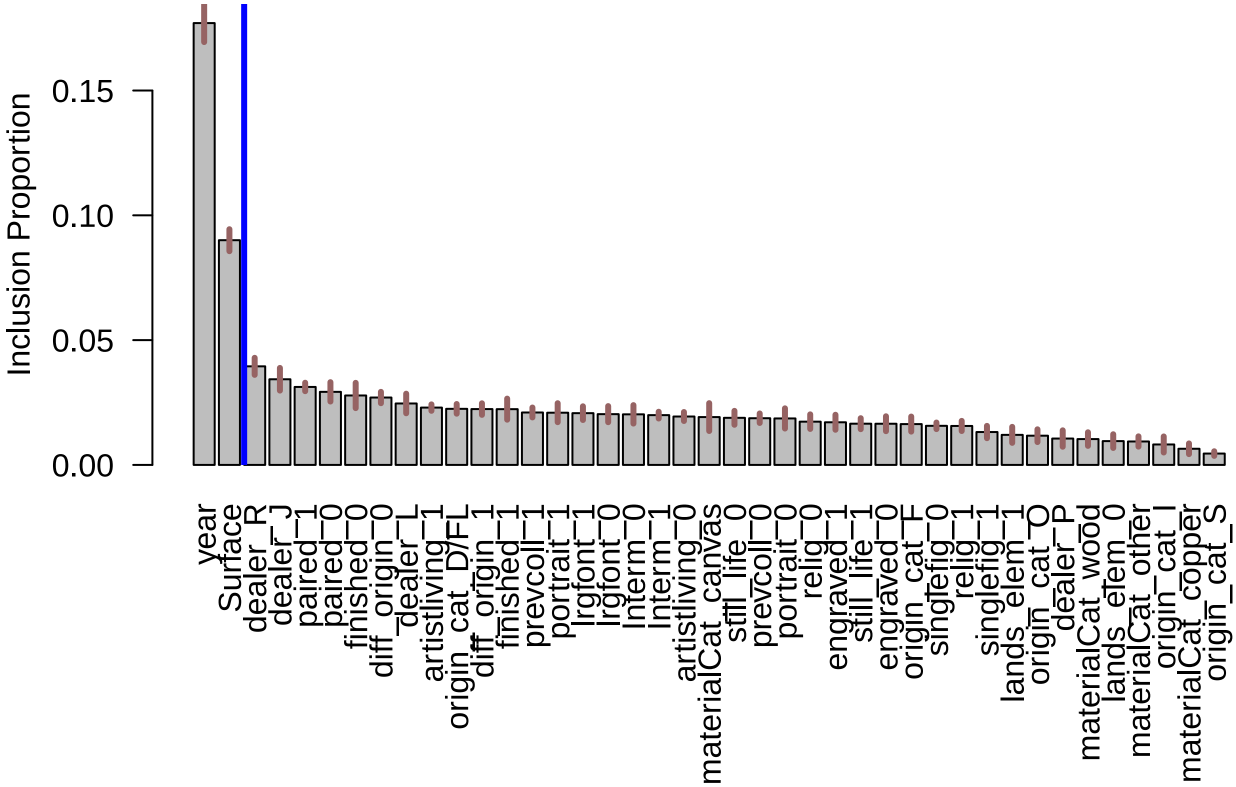
Development of the final model

To make our model more complex, we fitted both linear models such as BMA and nonlinear models including GAM, Random Forest and BART to the paintings data set. Based on the model evaluation, BART performed the best and thus we chose BART as our final model.

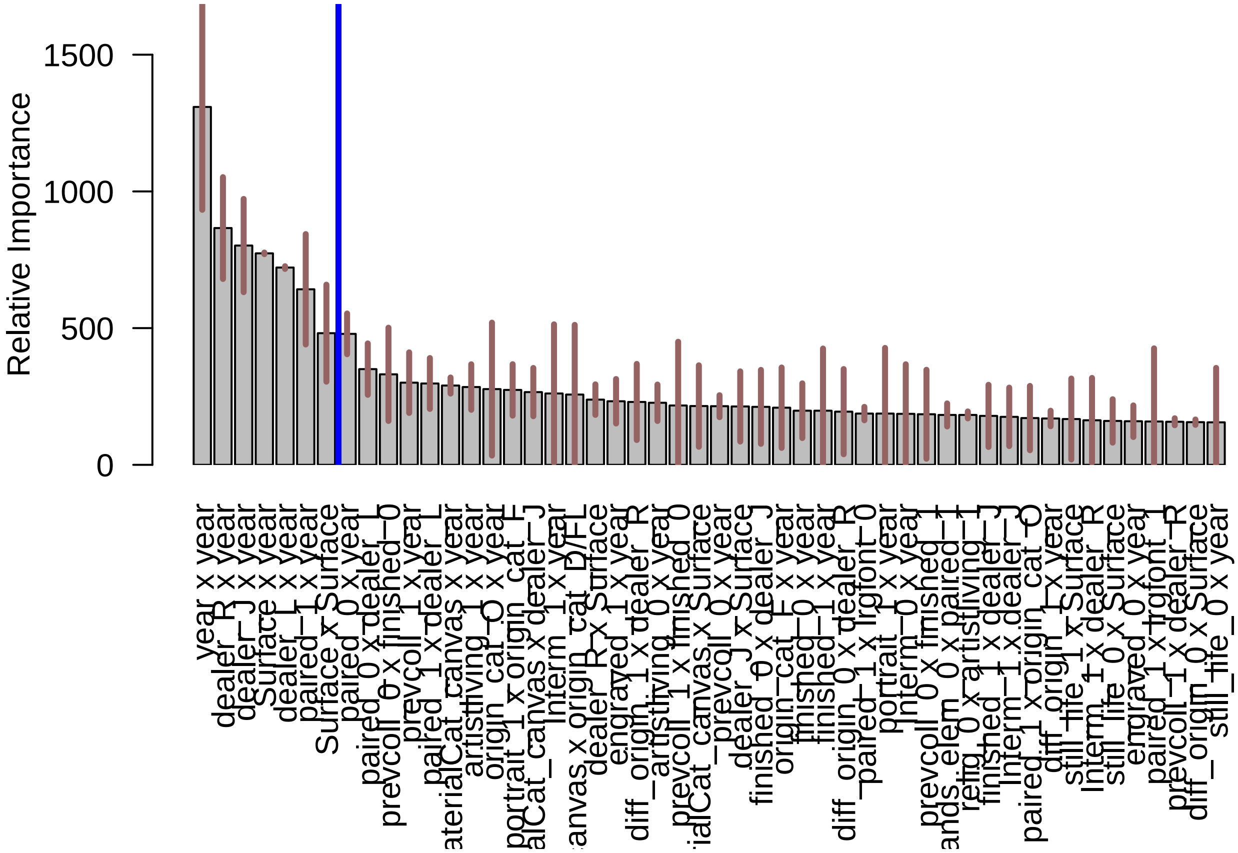
Final model

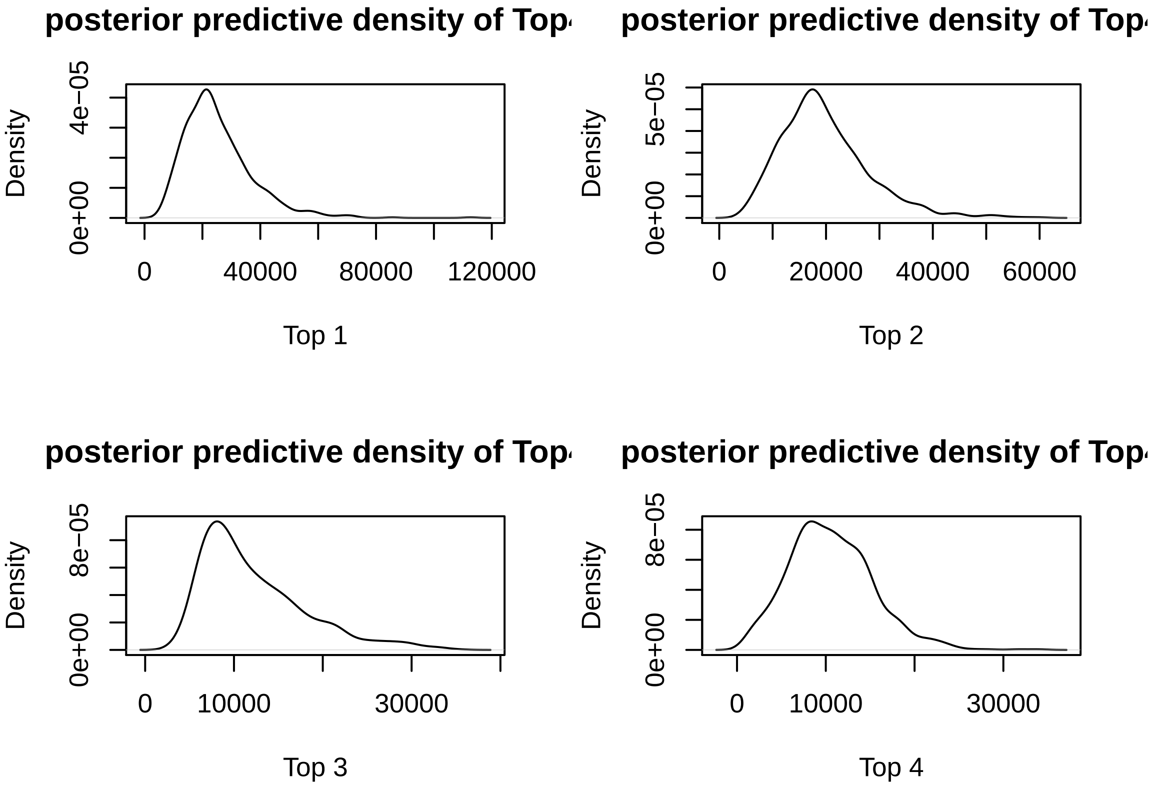
We provide plots of importance of variables and interactions for BART model since we cannot generate a summary table of coefficients like other linear models. Even though we can not present exact coefficients of our predictor variables and interactions, we could find their relative importance in our model. As we can see above, numerical variables; `Surface` and `year` have been included more frequently than other factor variables. In addition, two-way interaction terms including numerical variables are more frequent than other two-way interaction terms. The other thing we could find through above plots was that variance in interaction terms' including probability which indicated by red line on each bar were much larger than each individual variable.

```
## .....
```



```
## .
```

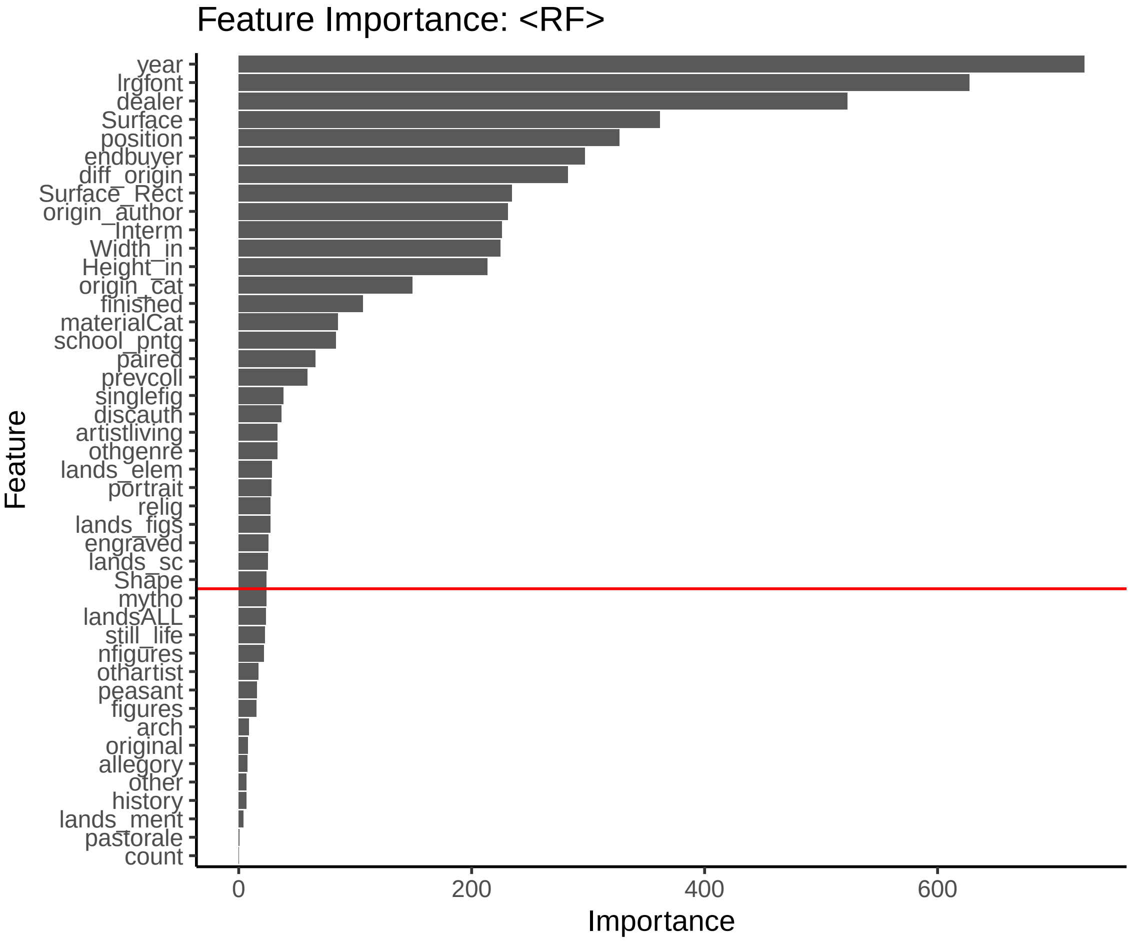




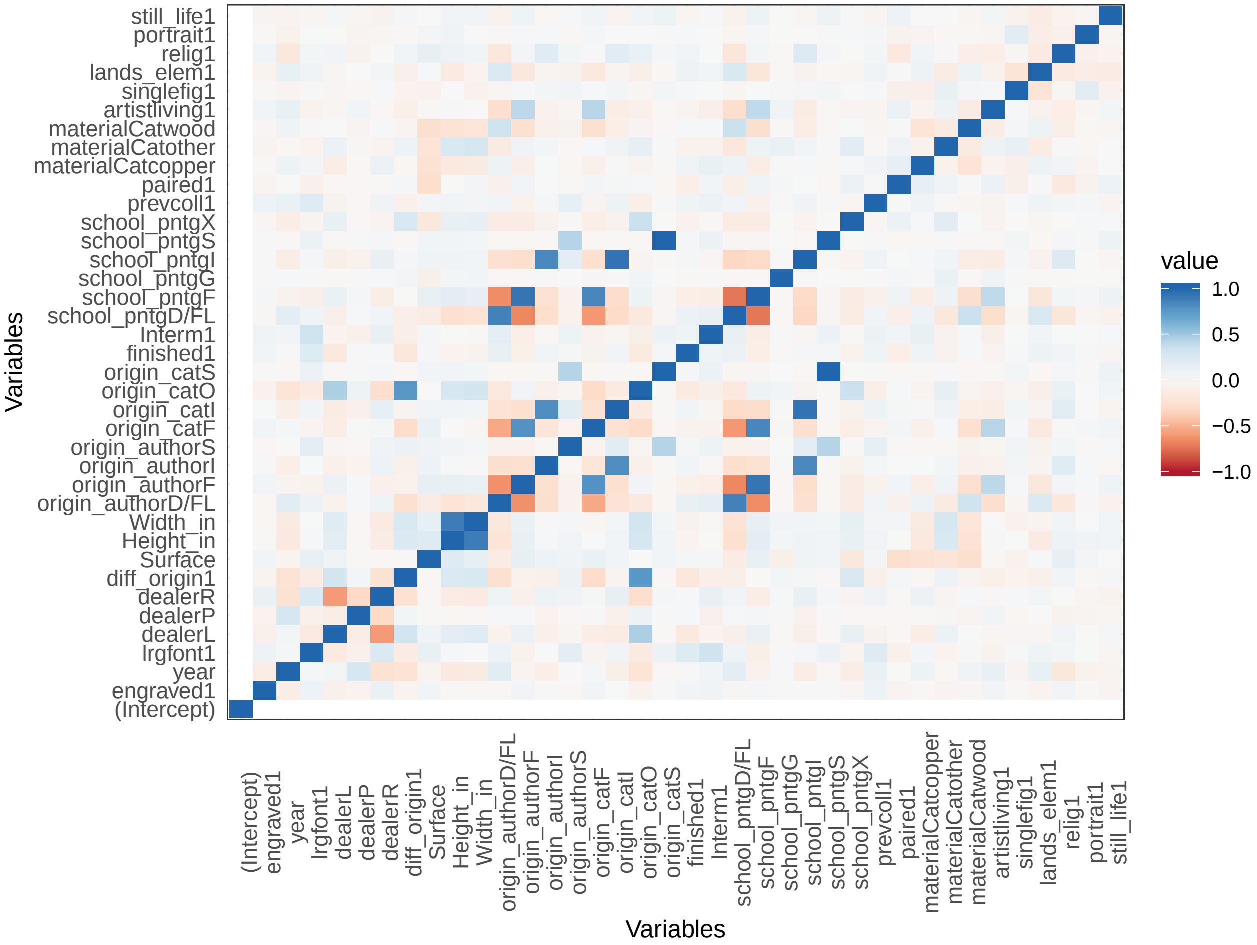
In addition to variable importance, we could find posterior predictive distributions of our test set which make it possible to quantify uncertainty. When we consider the most expensive paintings in train sets are 29000, `25800`, `20000`, and `17535`, the above posterior predictive distributions seem to capture the features in our dataset.

Variable selection

To improve our model in Part I and make our model more complex, we decided to include more variables in our model based on the variables we selected from EDA. Therefore, we tried systematic methods to select variables that we use in final model at first. Among 44 candidate variables, we arbitrarily choose top 28 important variables whose importance was measured by Random Forest.



To reflect the result from EDA, we included `still_life` and `finished`. However, if we use too many variables, there are potential problems such as unstability of design matrix or overfitting to training set. Thus, we excluded similar variables to prevent multicollinearity; `material`, `winningbidder`, `endbuyer`,`Surface_Rect`, `land_sc`, and `land_figs`. As a result, we had 26 candidate variables. We examine collinearity further to make our model stable.

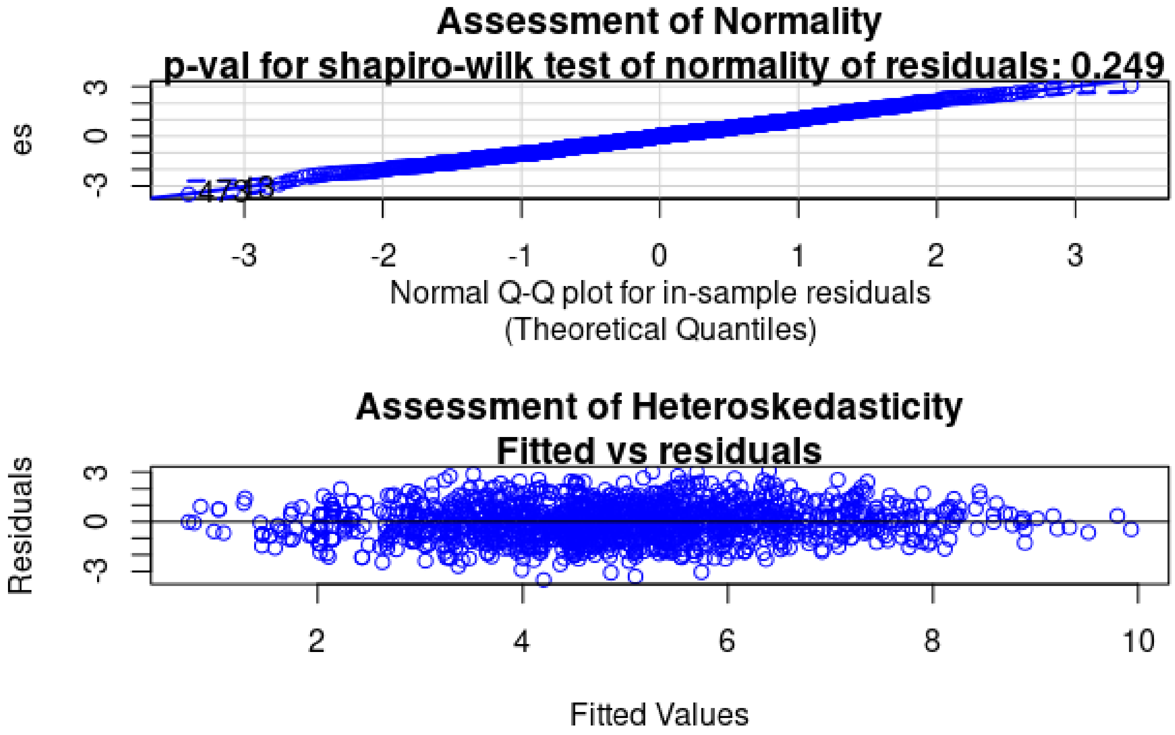


In above plot which indicates correlation between predictor variables, we could find `school_pntg`, `origin_cat` and `origin_author` are correlated each other closely. Moreover, `Height_in`, `Width_in`, and `Surface` are also correlated closely. Therefore, we decided to exclude simliar variables `school_pntg` and `origin_author` and keep `origin_cat`. Among `Surface`, `Width_in`, and `Height_in`, we decided to keep `Surface` because we thought that `Surface` can represent the others well. As a result, selected variables for final models are shown below:

Table 3: selected variable by RF, EDA, and correlation check

engraved	origin_cat	artistliving
year	finished	singlefig
lrgfont	Interm	lands_elem
dealer	prevcoll	relig
diff_origin	paired	portrait
Surface	materialCat	still_life

Residual



From the normal Q-Q plot, we can see that all residuals were perfectly followed a straight line indicated normality. Also, from the fitted vs. residuals plot, all residuals were equally spread around the horizontal line indicated linearity. Thus, our model met the general assumptions.

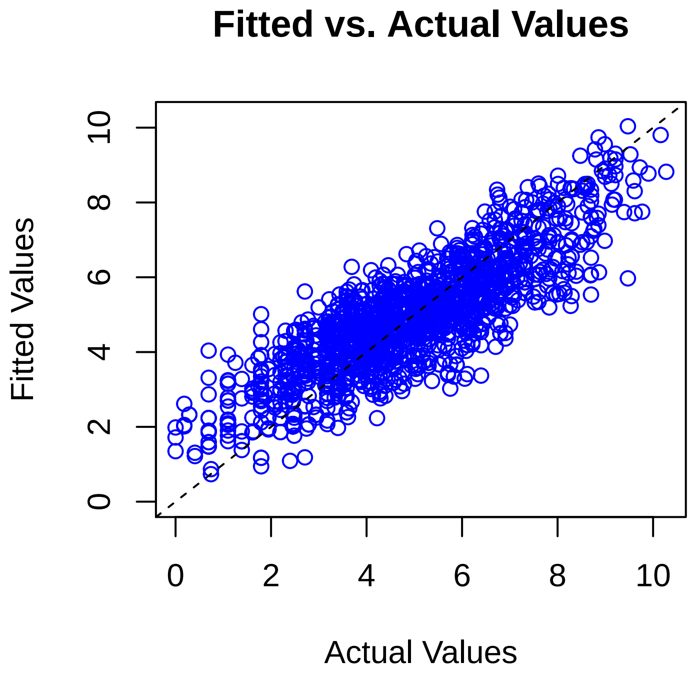
Discussion of how prediction intervals obtained

We used the function `calc_prediction_intervals` in the package `bartMachine` to constrcut the 95% confidence interval for our model. This function returns a matrix of the lower and upper bounds of the prediction intervals for each observation in the test data. We then exponentiated the prediction values, the upper and lower bound values back to its original units and putted into the appropriate data frame format for evaluation.

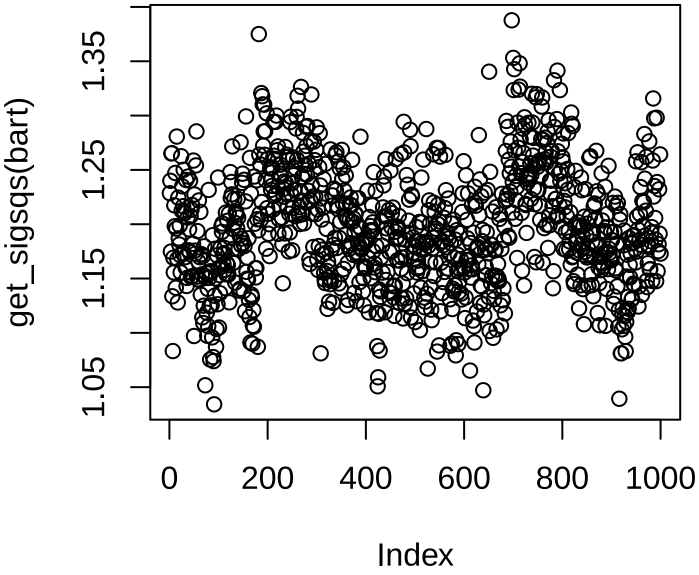
Assessment of the final model

Model Evaluation

The plot of actual versus fitted values indicates that the fitted values from our BART model is reasonably close to the actual responses on the scale of log(price). Notice that the fitted values are slightly higher than the actual values for lower prices, and slightly lower than the actual values for higher prices, suggesting that our BART model tends to over-predict the prices of the paintings when the actual prices are relatively low, and under-predict the prices the paintings when the actual prices are very high.

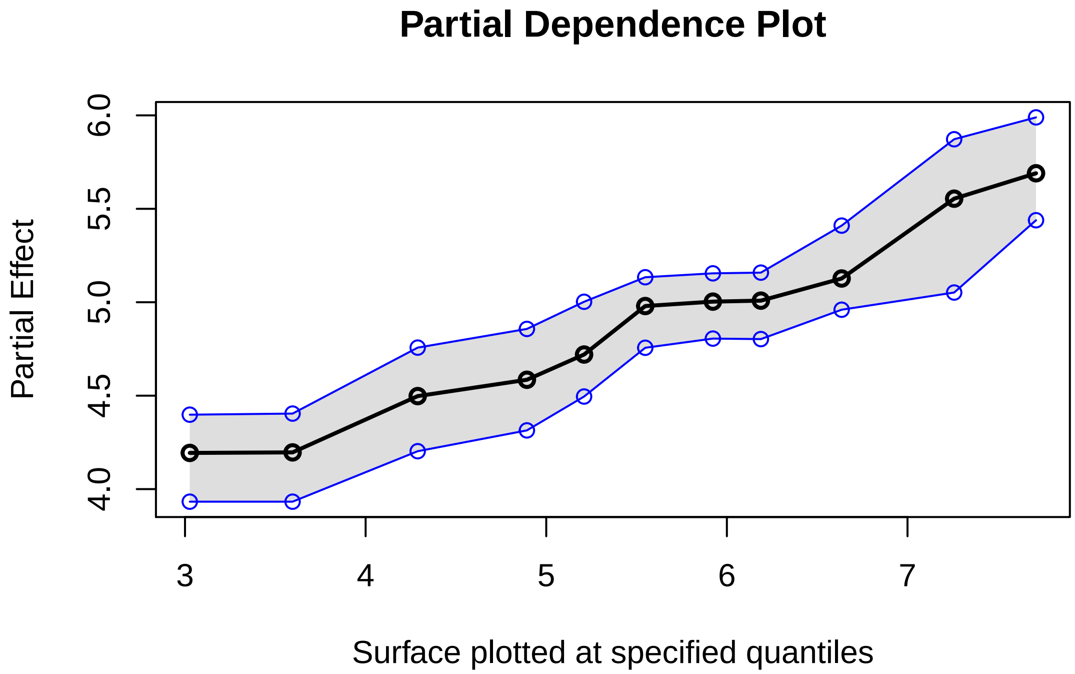


The following plot is a histogram of the posterior estimates of the error variance from the Gibbs samples. The y-axis is the posterior estimate of the error variance. Notice that the error estimates are quite small, which are desirable.



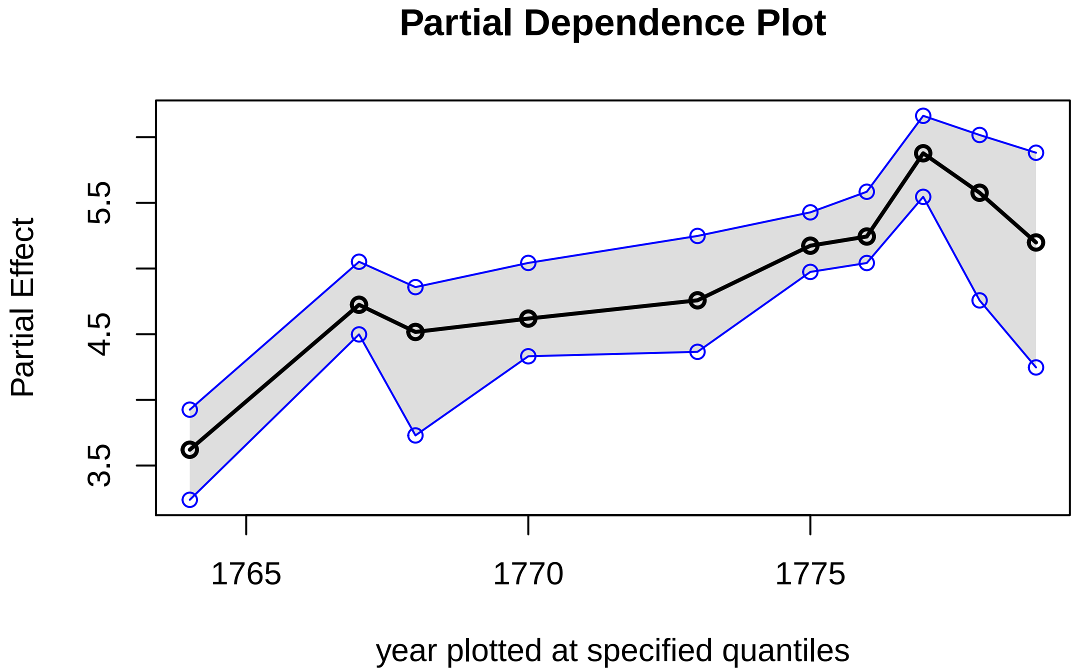
We only explored the partial dependences for the 2 numerical predictors `Surface` and `year`. The following partial dependence plot shows the partial marginal contribution of `log(Surface)`. Since the partial effects do not include 0, and increase with an increasing surface area, the plot shows that `Surface` contributes quite a lot to the prices of paintings.

```
## ...........
```



The partial dependence plot below shows the partial marginal contribution of `year`. Since the partial effects do not include 0, and increase as the year increases, the plot shows that `year` also contributes quite a lot to the prices of paintings.

```
## ..........
```



Model Testing

BART combines the advantage of boosting by decreasing the residuals from the previous trees fitted, and increases the size of tree without suffering from overfitting. The overall performance of our BART model is good, as can be seen from the following summary table:

	Bias	Coverage	maxDeviation	MeanAbsDeviation	RMSE
BART	177.967	0.964	12708.83	456.433	1237.631

We achieved a very low bias and RMSE, suggesting that our BART model can make very accurate predictions. The bias is low, meaning that error introduced by approximating the model by our BART model is small. The RMSE is also small, which indicates that the variances of our residuals are also small. The coverage of our model is 0.964, which achieves a quite high predictive coverage. Both the low RMSE and high coverage indicate that our BART model is adequate. In addition to BART, we also tested other models including OLS linear regression model, BMA, GAM and Random Forest. We included a summary table of the model performances in the later sections of our report.

Model Result

We predicted the price of paintings based on the validation data, and ranked predicted price from the highest to the lowest and summarized the information of top 10 paintings according to the variable importance of the bart model. The results are shown in the table below:

year	dealer	paired	diff_origin	Predicted Price
1777	R	0	0	16997.096
1776	R	0	0	14306.575
1777	R	0	0	11890.768
1777	R	0	0	10403.958
1777	R	0	0	9419.412
1777	R	0	0	8799.996
1777	R	1	0	8779.777
1767	R	0	0	7607.519
1777	R	0	0	7045.205
1777	R	0	0	6370.088

It's found that most valueable paintings are sold in year 1777. All top 10 paintings are involved with dealer R. Almost all paintings are sold or suggested as not a pairing for another. In addition, the origins of all paintings based on nationality of artists are the same the origins of paintings based on dealers' classification in the catalogue. We don't observe significant trend in top 10 paintings, but it might contribute a lot to the price for the rest of paintings.

The highest predicted price is around 14000. It's found that year, surface, dealer, paired, diff_origin are important variables to predict the price of paintings.

Conclusion

Summary of Results

	Bias	Coverage	maxDeviation	MeanAbsDeviation	RMSE
BART	197.946	0.964	12708.83	440.461	1219.663
Random Forest	206.970	0.947	13774.25	472.354	1252.426
BMA	218.632	0.945	13871.02	474.235	1263.940
OLS	335.079	0.899	15610.29	449.551	1323.228
GAM	130.507	0.237	26822.44	534.276	1772.144

To find the factors that would drove the prices of paintings, we cleaned the dataset and imputed for missing values. We started from a simple linear model that only includes predictos selected according to the EDA to give us some ideas of potential relationship between the price and other features. We then fit complex models which are linear model, BMA, BART, Random Forest and GAM with predictors selected from results of Random Forest and EDA for a better performance. The tested performance is listed in the above table. By taking account all information in the table, we choose BART as our final model.

Even if we cannot interprete coeffcients of variables in the model like what we did in linear models, we still find some important features that could make prices of paintings high. It's found that numerical variables including `Surface` and `year` have significant effects on prices. Paintings sold in later years tend to have a higher price. Large surface area could improve price, and the effect is not that obvious and significant. We also find if dealer R is involved, the price would be higher. In addition, paintings with higher price tend to have the same origin based on nationality of artists as the origin based on dealers' classification in the catalogue. Also, it's better not a be paired with another painting. In conclusion, to get a painting with higher price, people should choose a painting that's sold in later years, with comparatively bigger surface area and not be paired with others. It's also better to have dealer R, and origins of authors and dealer's catalogues are the same. Among all of features, `year` is the most important one.

Discussion of thing learned

- **Clean the data and use "mice" to impute for missing values?** Cleaning data is always the first step of statistical analysis. It took us a long time to figure out what each variable represents, how to code variables in the desired type including removing redundant information such as strange symbols and relevel factors as needed. A tidy dataset is always the foundation of a good model analysis. Since the dataset in the real life is often messy or not in the format we prefer, we need to spend more time making the data frame as clean and efficient as possible. One common problem that we always encounter is that there're missing values in several variables. At the beginning, we considerd training NA as a new level for factor variables. For numeric variables, we tried to use the median to replace NAs of that variable. This method could save time. However, it doesn't make sense when we interprete our results for this new level, since the data doesn't actually have this category. The final method we adopted is to use "mice" package to impute missing values according to other columns. In this way, we neither generate new levels nor randomly assign values. All imputed values are based on our existing information, which would keep the bias as low as possible.
- **How to effectively choose variables in the model?** Exploratory data analysis is the first step to know the relationship between variables. The first set of predictors we chose for the simple linear model comes from EDA. We could use AIC or BIC to select important interactions after we have some desired predictors. For further analysis, we could use complex models including BMA and random forest that could generate the marginal inclusion probability or the importance plot showing what variables could be considered as important. After comparing and combing predictors selected by different methods, we could refine the set of predictors that would be very likely to generate a good model.
- **Effects of collinearity between variables** We should remove some variables that're obviously dependent. For example, `Width_in` and `Surface`. We might need to remove one of them. For some variables that are hard to see correlations directly, we need to draw some plots to find if there're some related predictors. The performance of models would be significantly improved if all predictors in the final model are independent.
- **Advantages of BART model** Although it's difficult to interpret the BART model, BART could handle overfitting. In general, it seems that tree models might perform well in this project.

Recommendations and suggestions

If we have more time, we would try other models such as Ridge and Lasso to see their performances. Moreover, we would try to learn more about package bartMachine and try to tune the BART model to get a better result. We would also try to improve the random forest model and try bagging and boosting to see if these models could perform even better. Finally, we don't use the predictor "author" in the dataset since there're too many authors, but authors could be intuitively thought as closely related with the price of paintings. Thus, there might be a need to clean variable "author", regroup it and include it in the analysis.