Part-II Writeup

Team: SparkR

2017/12/7

Introduction

Late 18th century Paris has witnessed countless number of auctions on fine arts and paintings from countries across Europe. Our project utilizes this dataset from auction in 18th century Paris with details of paintings, auctions, dealers and other third parties involved.

We hope to build a model to predict the log price fetched at auction. Since our dataset contains 59 variables, we would want to perform variable selection to build a parsimonious model with good performance in critiria such as RMSE, etc. For this part, since we are not restricted to OLS, we did more exploration into models we have learned this semester, e.g. random forest, boosting, bayesian additive regression, tree.

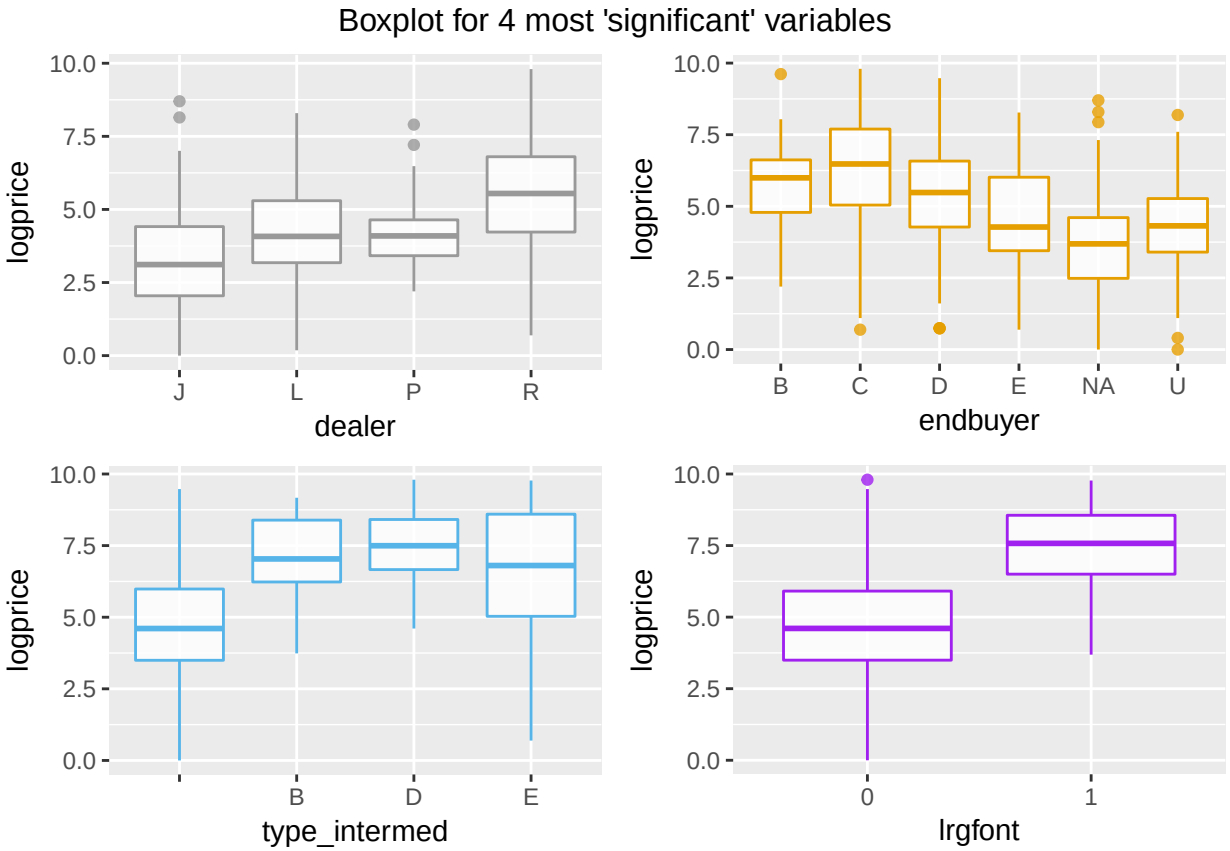
In addition to that, some variables have a lot of missing values and typos. We would need to perform some sort of missing imputation, e.g. assume MCAR and use MICE package. For typos, we would recode them to appropriate values. During predicting with validation data, we found some new levels in some variables, e.g. material. So we modified our data cleaning process so our dataframe could be used for various models.

From the correlation martices, we find some highly-correlated variables, so we decide to not include some of the variables in our model to aviod multicollinearity. Algorithm such as gradient boosting handles highly correlated predictors very well, so models will be robust.

Exploratory data analysis

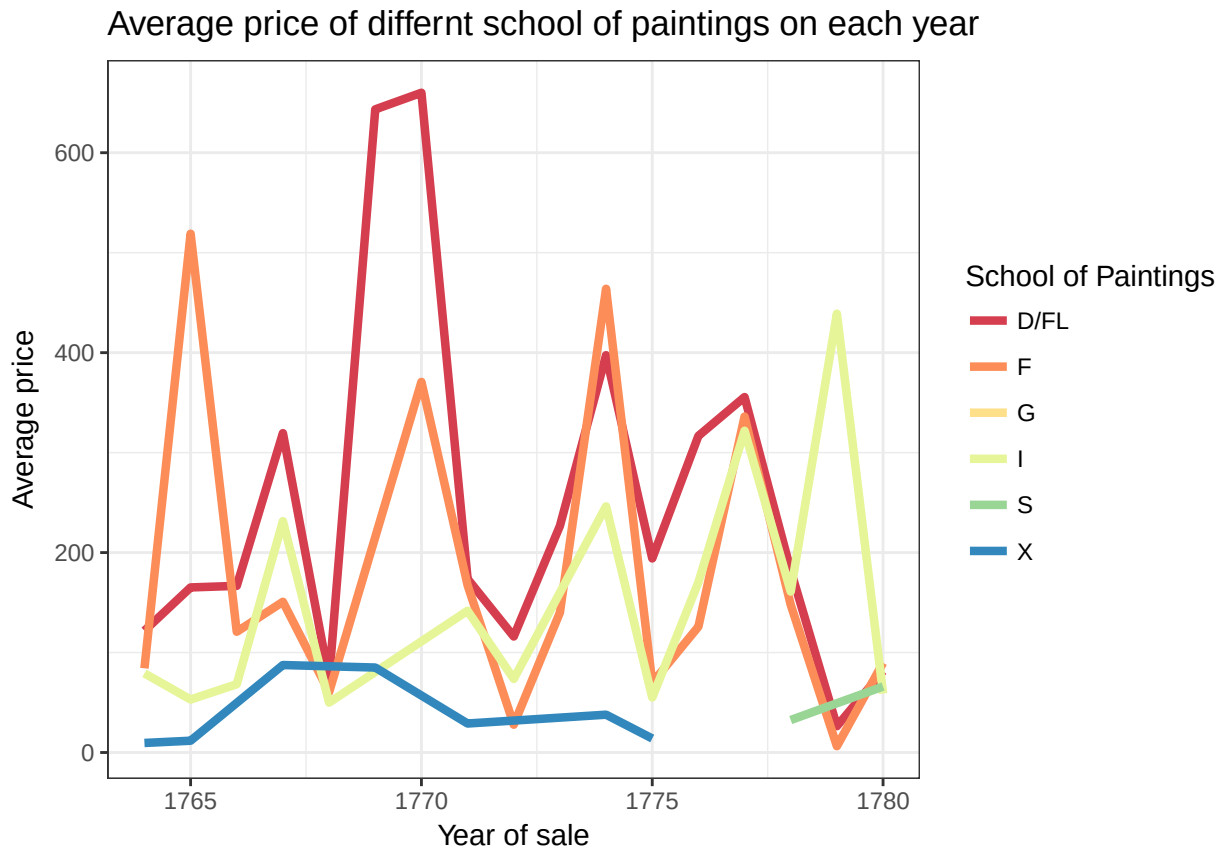
For this part, we intially have found some interaction between predictors from graph perspective; however, in our further model selection process and trials and error.

First, we choose to plot boxplots for the variables that are highly correlated with response `logprice`



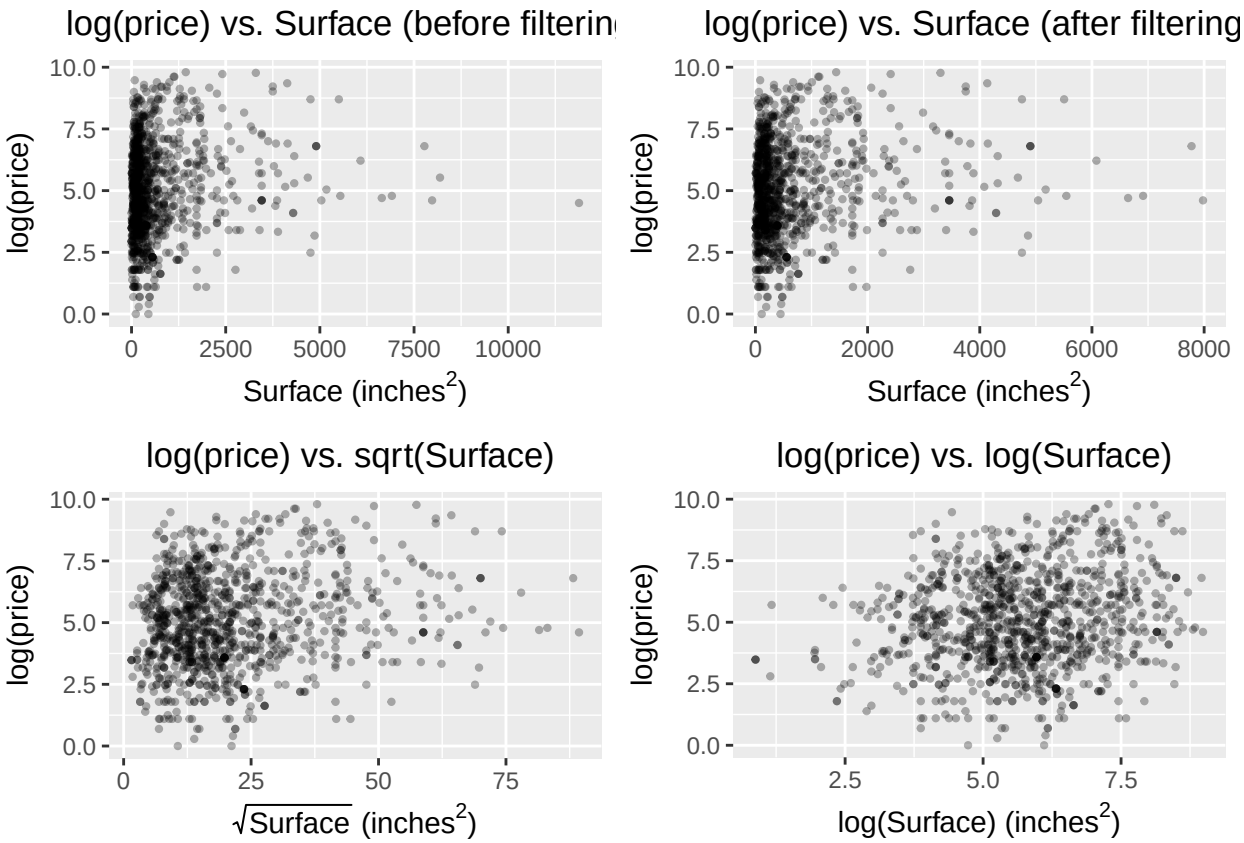
As we can see, for these four factor variables, each level has differernt means and distributions across, for examle, doing business with dealer R may not be a good deal, since they tend to have higher price, but it could also be the case that they tend to sell higher value painting; also, collectors tend bid higher price compared with other end buyers; from type of inttype of intermediary perspective, experts tend to have higher price in mind,(becasue they know the true intrinsic value?), lastly, if dealer have more information about the paintings, the final price will be higher as well. Therefore, it is likely human factors(especially type of dealer) play an important role on deciding the final bidding price, and we will expect they will "survive" after model selection process.

Second, we try to find some interaction effect between predictors using tree models.



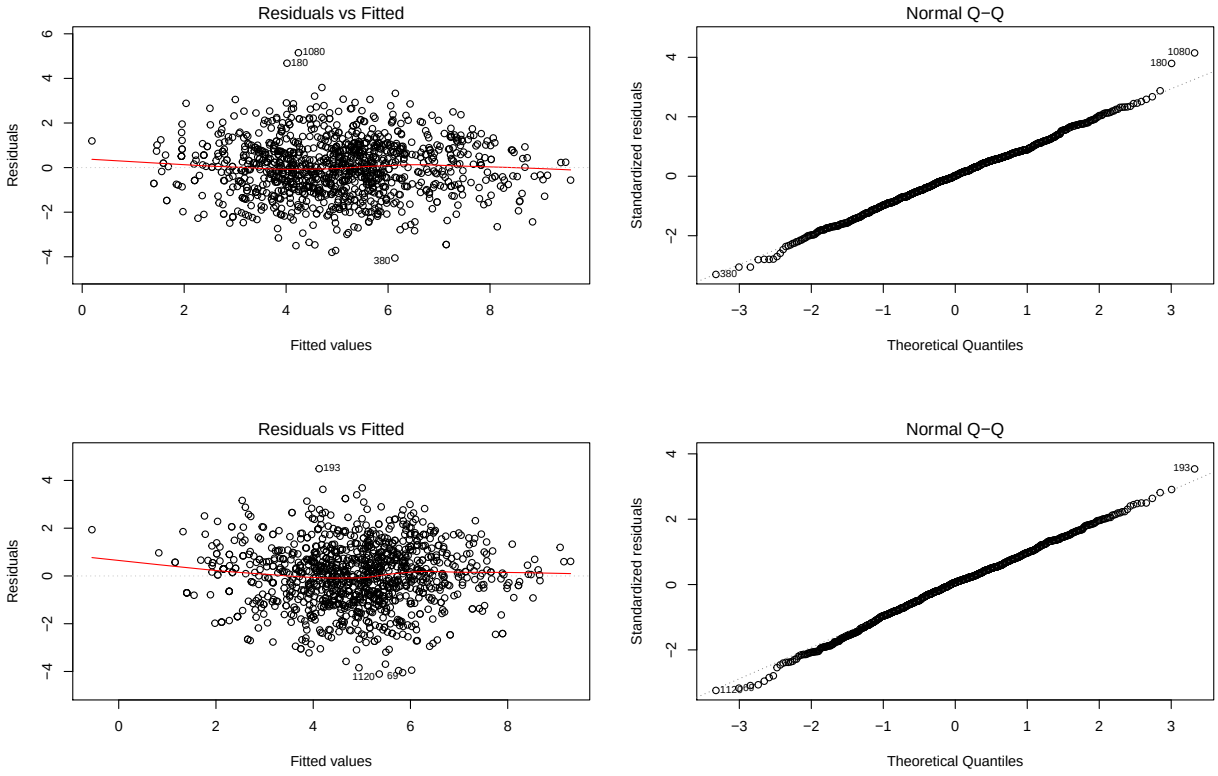
Since when plotting average logprice, all the line lie between each other, we decide to plot price in original scale instead. As you can see from the plot, Dutch/Flemish and French school of paintings tend to have a higher average price in the year between 1765 and 1778, with the maximum across the group, Dutch/Flemish has the average price over 600 dollars for the paintings in the year around 1770(possibly becasue of the decease of some well-known artists);Even though Italian has lowever in the earlier year, they do have higher price in post-1777 era. With all these observations, it suggests there will be "some" interaction effect between `school_pntg` and `year` that affect the overall price of paintings.

Third, we dive into predictors transformation:



At a first galance, we do not observe any significant linear relationship between the response: "log(price)" and the predictor: "Surface". After we remove some outliers and zeroes(0 surface is not meaningful), we do not achieve much improvement. Therefore, we decide to use transformations on "Surface". First, we take the squre root on "Surface". As we can see, the points are less concentrated and more normally distributed after the tansformation. When using log transformation, we have the distribution of log(Surface) very close to normal distribution. Under log transformation, we still cannot see any obvious linear relationship between the response and the predictor. However, when we perform variable selection under Bayesian information criterion, we would have "Surface", log transformed, included in the final model with a positive coefficient and a relatively small p-value. By intuition, the inclusion of Surface in the model makes sense because paintings with larger surface would probably correpond to more amount of work and more painting materials used and therefore higher prices.

Fourth, we discover serious issue in our previous model:



As you might remember from the part I, the upper half is the diagnostic plot using our proposed model in trainning data, even thogh from residuals vs fitted plot, residuals are around 0 and variance about the same in all range of expected value; from normal Q-Q plot, there is no indication of violation of normality. However, the lower part reveal different information, as we use the same regression model in the test data, there is a significant problem of overfitting, since all the residuals are bascially lying arond zero, the residuals show no sign of normally distributed, as the model can "perfectly" all the data point in test data. Hence, in our future model, we need to use different approach to predict validation data.

Discussion of preliminary model Part I

The overall performance for test data with our PartI model is somewhat misguiding, with the RMSE is 1273, and the coverage is about 93%, the overall performance is masked by the fact that the problem of overfitting(only add predictors that match data behevior in testset), since compared with RMSE in the test data, RMSE in the training set is about 2000; therefore, in order to robustfully predict validation data, we need to come up with a different predictors(or models) in order to corporate differenct instances of behaviors in each data set. Futher development could be use the predictors selected by step function with AIC critiria as well as BMA that can combine different models and lower the variances.

Development of the final model

- Data Set Modification:

As mentioned in the our previous exploratory data analysis part, although we had relatively low testing rmse for the preliminary model we developed in part I, the model actually has an overfitting problem. The model would produce high rmse for the training data set and the diagnosis plots also indicate that the model fits

the test data set too closely as all the residuals are close to zero. Therefore, as we were developing our final model, we wanted to balance the performances of our models on the training data set and the testing data set. Although the accuracy of our predictions for test data set is important, we still want to avoid overfitting problem. We started with a full model with all the available predictors selected with some modification, filtering predictors for redundancy, multicollinearity and adding some factor variables indicating if a painting has certain features. For redundant predictors such as "mat", "material" and "materialCat", since they all contain the same information, we only kept "mat" and dropped the other two predictors. For predictors with high multicollinearity with each other, we filtered those less important variables. For example, for "Height_in", "Width_in" and "Surface_Rect", we kept "Surface_Rect" to represent the surface area of the paintings and dropped the other two predictors. We also added new predictors such as "SurfaceNA" to indicate if the surface of a painting is specifies at the auction. Besides these modifications of the training data set, we also needed to handle the predictors having new levels in the testing and validation data sets. For such predictors, since usually the number of occurrences of their elements with new levels are small, for most of the times, we included them to an already existing or newly created category: "other". As mentioned in Part I, we also converted the data types of predictors to appropriate types and perform some transformations on predictors.

- Predictor Selection

After we have completed our data set modification, we tried various models with all the predictors we have: LASSO, BMA, random forest and BART. Among these models, only LASSO would preform some sort of free variable selection and therefore control the number of predictors while as indicated by the lectures: "BART is similar to Boosting in that the mean function is a sum of trees, but uses a Bayesian approach to control complexity. Trees can be of different sizes and the number of trees can be large without overfitting". Therefore, in order to control the previous overfitting problems, we want to perform some varaible selection based on AIC, BIC or other criteria.

Under AIC(forward), we have the following predictors:
position,dealer , year , diff_origin ,origin_author, origin_cat, endbuyer , type_intermed , Surface , nfigures , engraved , prevcoll , artistliving,original,othartist, paired , finished , lrgfont, othgenre , portrait , still_life , figures,relig,landsALL, discauth , school_pntg_recode , SurfaceNA,lands_sc, singlefig, history,allegory, pastorale, other, mat_recode, shape_recode

Under BMA(posterior probability larger than 0.5), we have the following predictors:
dealer , year , origin_author , origin_cat , diff_origin , endbuyer , type_intermed , Surface , nfigures , engraved , prevcoll , paired , finished , lrgfont , othgenre , portrait , still_life , discauth , school_pntg_recode , SurfaceNA , shape_recode

After comparing the training(calculated locally) and testing RMSE(from the leaderboard) for each model mentioned above with full predictors, predictors under AIC and predictors under BIC, we find that the predictors selected under AIC has the best overall performance.

- Model Selection

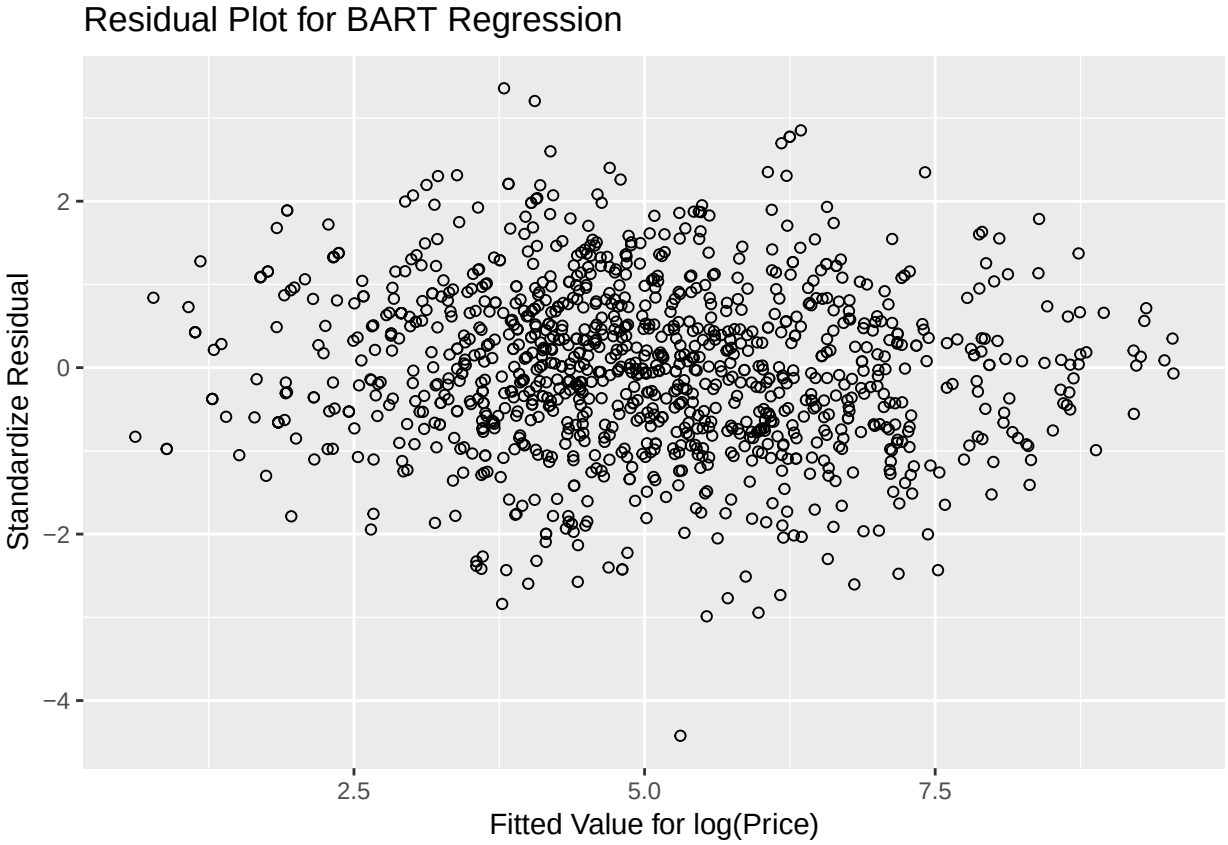
After we selected the predictors, we would like to compare the models, we used the training and testing rmse summarized in the following table 1:

	Training RMSE	Testing RMSE
LASSO	1452.715	1677.637
BMA	1522.493	1527.934
BART	1360.145	1423.688
Random Forest	1617.477	1487.458

Based on the table(testing RMSEs are form the leaderboard), among the models, BART and Random Forest has relatively low testing RMSE below 1500. However, Random Forest is possibly overfitting the test data with relatively high training RMSE while BART has the lowest training RMSE. Moreover, BART

performs better than LASSO and BMA in both training and testing RMSE. Thus, we selected BART with the predictors selected previously as our final model.

- Residual Plot and Discussion



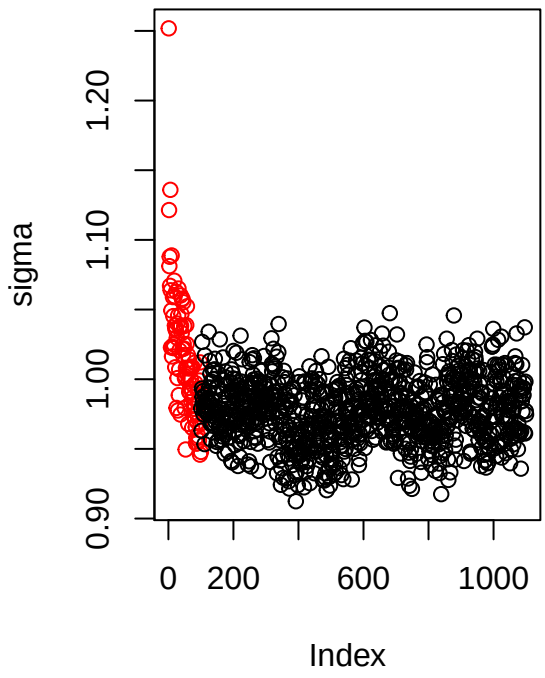
Discussion:from the Residuals vs. Fitted diagram, we can see no distinct pattern indicating non-linear relationship. Residuals are generally equally distributed around a horizontal line. Therefore, the linear relationship assumption appears to be met.

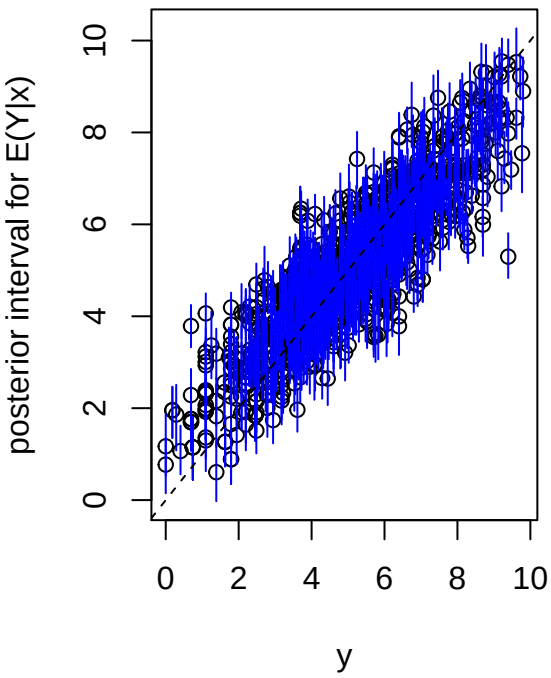
- How Prediction Intervals Obtained

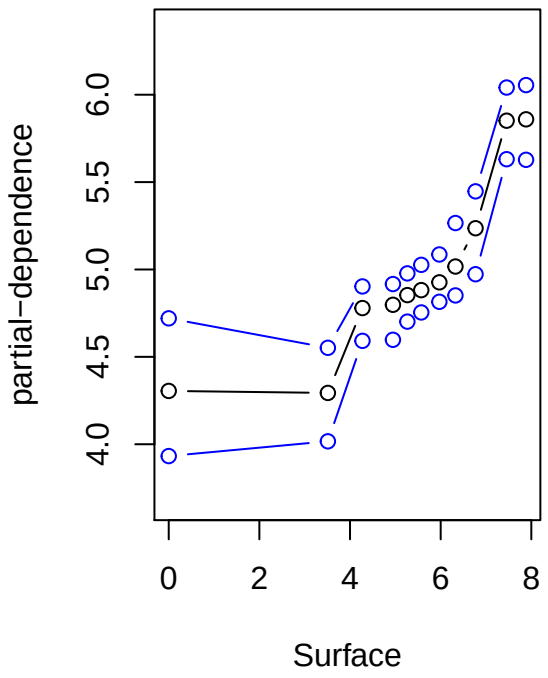
Note that for the confidence interval, we used "bartMachine". Since "BayesTree" package has a built-in predict function, since it's uses bootstrap method and running long iterations is painfully slow. But small sample size could be inappropirate to construct a prediction interval. Hence we uses "bartMachine" which parallelize its calculation and runs much faster.

Assessment of the final model

- Model evaluation:


sigma
0.90
1.00
1.10
1.20
0
200
600
1000
Index


posterior interval for E(Y|x)
0
2
4
6
8
10
0
2
4
6
8
10
y



From the first plot, it shows the uncertanty of the predictions, the uncertainty around the bounday of y(ie. 0 and 10) is larger than the one in the middle. Since `pdbart` function is computationally intensive, we only plot one continious variable "Surface"; For its partial dependence plot, it shows parital marginal contribution of "Surface", since the plot doesn't include 0, and the dependence increase the size of surface increase, it shows that this predictors contribute a lot.

- Model testing

As mentioned above, we have tested other models as well, even though Bart lack of interpretation ability, its combine advantage of boosting(decrease residual of previous tree by) and can increase size of the tree without overfitting. As the similar case with LASSO and random forest, it is hard to obtain confidence interval, although the prediction come with the form of matrix, with 1131 observation and 1000 instances for each observation, if we use confint with c(2.5%, 97.5%), the coverage show us under fit, so we decide to use `BartMachine` instead as suggest in the lecture. Granted, there is no variable selection within Bart process and it provide no interpretation of varible in the end, with the lower RMSE across the board, we finally choose Bart after all.

- Model result

The results for the top 10 valued paintings in the validation data are shown below:

dealer	endbuyer	lrgfont	price_pred
R	C	1	26411.374
R	C	1	26027.726
R	C	1	16488.834
R	C	1	14044.620
R	D	1	13790.643
R	C	1	12997.283

dealer	endbuyer	lrgfont	price_pred
R	C	1	10953.605
R	C	1	10287.039
R	C	1	9776.950
R	C	1	9359.128

As the result shows, all top 10 price painting are involved with Dealer R, and the paintings are end up with collecter and dealer, futhemore, dealer devote an additional paragraph when introduce the paintings; all these results show consistency with our assumption in partI EDA anlysis.

Conclusion

Summary of Results:

The performances of the models in Part I and Part II on the testing and training data sets are summarized in the following table.

	Training RMSE	Testing RMSE
LASSO	1452.715	1677.637
BMA	1522.493	1527.934
BART	1360.145	1423.688
Random Forest	1617.477	1487.458
OLS	2145.830	1272.900

As we can see from the table above, for our OLS model in Part I, although we achieved the lowest testing RMSE, we found that the OLS model has overfitted the test data set and therefore the resulting training RMSE is extremly high. For the OLS model, in the summary, we have R-squared equal to 0.5899, which suggests that 58.99% of the variation has been explained by our predictors.

In Part II, we tried to resolve our problem by balancing the performance of our models on testing and training data sets. We tried LASSO, BMA, Random Forest and BART models with predictors selected by AIC, BIC and posterior probability in BMA. It turns out that the BART model with the predictors selected under AIC gave us the overall lowest and training and testing RMSE as well as relative balanced performance over testing and training data sets.

Things learned:

- The first thing we learned is data cleaning. It costed most of our time during first part of our project. Most datasets we will see in real life could be somewhat messy and need a lot of preprocessing before we actually start modeling. Since our train, test, validation sets have the same structure. We coded a function to clean up our dataset to streamline this process. In addition, some categorical variables in test/validation sets have more levels, which we needed to account for and update our cleaning process.
- Secondly, we learned the importance of not overfitting test data. During the first part, we tried hard to get on top ranks on the leaderboard and ended up with test RMSE of around 1270 but train RMSE with more than 2000. It's clear that we overfitted the test data and we do not expect such a model will perform well on the validation data. Later we tried other models such as LASSO, Bayesian Model Averaging, trees, random forest, boosting, extreme gradient boosting and Bayesian Additive Regression Tree. Some models run into the issue of overfitting part of the dataset. We could examine this by performing cross validation to see which model has the most balanced RMSE for train and test data. However, some models are more robust for overfitting, such as extreme gradient boosting and BART. XGBoost has a built in parameter(gamma) to control overfitting. And as mentioned in lecture, BART

handles overfitting very well, and as for our datasets, BART has the most balanced RMSE(and it's low).

- Last, we learned that compuation time could be a huge problem. Since the dataset have more than 40 variables, fitting models such as XGBoost is not as easy as fitting a linear model. Also, since tree/boosting methods do not have inherent confidence/prediction intervals, we need to take bootstrap samples and it could take a long time. At last we learned a package "bartMachine" which uses Java for parallel processing.
- If we have more time: we would explore more with bartMachine and add a cross-validation method, which could improve our results and make our model more robust. Unfortunately, our laptops have limited RAM and we might need to distribute the compuation on cloud service, such as AWS. In addtion, we did not use author as a predictor but intuitively, it could be an important predictor. Since author has too many levels, we could recode this variables according to how celebrated the author was.