Final Project Part II

Team_FP04

12/13/2019

Introduction

The 18th century is often refered to as "le siècle, des Lumières" (the century of the lights) in reference to the philosophers that emerged early on the 1700s leading the way towards the French Revolution. In addition to its important societal evolutions, the 18th century was also a major period for art in France and it is therefore of interest to understand painting trading during that period, especially before the French Revolution (1789).

The aim of our analysis is to explore the factors that drove painting prices in 18th century Paris. The painting prices will be predicted from auction price data between 1764-1780 containing information on the sale (seller/buyer), the artist, and other characteristics of the painting. This analysis will also allow us to assess which paintings were overvalued or undervalued.

In the first part of the analysis, We first explored the effects of potential predictors and their interactions on painting prices by conducting an Exploratory Data Analysis. This also allowed us to prepare the data for the next phase. We then built a linear model using stepwise regression method with akaike information criterion (AIC) and a training subset to select a robust model predicting the auction price (using the log transformation `logprice`). We finally validated our model on a test subset.

In the second phase of the analysis, we will revisit the linear model and assess its quality of the fit. Given the limitations of the linear model, the results might be improved by an optimization of the bias/variance trade-off. A more thorough EDA will also be considered. We will then explore more complex techniniques to produce a better fitting model that should have better out-of-sample predictions of auction prices. We will consider several different options, including Bayesin Model Aaveraging (BMA), random forests, boosting and Lasso, in order to find the model that performs best on the test data.

EDA

We start by exploring variables in the training dataset and understand their meaning. We first implement the required data pre-processing:

- transform empty string and "n/a" character to NA
- Delete duplicate rows
- Transform binary and charcter variables to factors
- Change `position` values not bounded from 0 to 1 to NA
- Reconcile `Shape` coding (ovale = oval, round = ronde)
- Group `authorstandard` and `winningbidder` into quartiles
- Impute the missing data

Imputation: We impute the mean of each quantitative column except Surface (heavily skewed right) for which we impute its median. For the binary and multiple level factor variables, we choose to impute the mode of each column.

we then see that some variables were used to classify each painting and therefore cannot be used in our analysis: `sale`, `lot`, `count`, `subject`, `author`, `subject`, `authorstyle`

We exclude `authorstandard` and `winningbidder` and keep their quartiles classification. In addition, since we are predicting `logprice`, we will not used variable `price`.

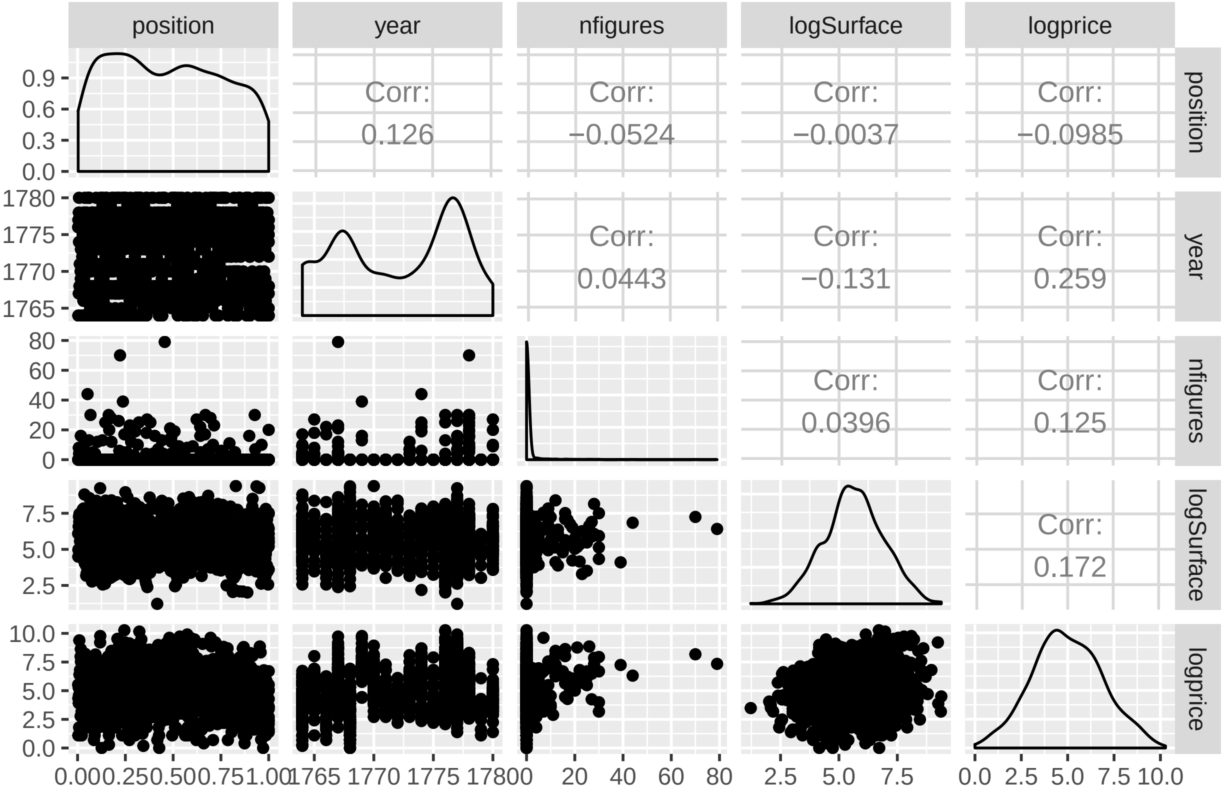
We can classify the remaining variables in two different ways, either by the way they are coded (quantitative, dumy, multiple level factors) or by the information they provide (i.e. sale, author, size & material or characteristics). First, we goup variables according to the way they are coded:

Quantitative Variables

`position`, `year`, `Height_in`, `Width_in`, `Surface_Rect`, `Diam_in`, `Surface_Rnd`, `Surface`, `nfigures`

We decide to classify `year` as numeric in the displayed EDA as it is spread around more than twenty years. Using our intuition, we choose to only use `Surface` and drop `Height_in`, `Width_in`, `Surface_Rect`, `Diam_in` and `Surface_Rnd` as they are extremely correlated and would not necessarily bring any additional information. We might want to later investigate the relationship of `Surface` with other variables such as `Shape` or the type of material. We use a scatterplot matrix to investigate the relationship between these quantitative variables and `logprice` (FIG 1).

FIG 1: Relationship of relevant quantitative variables

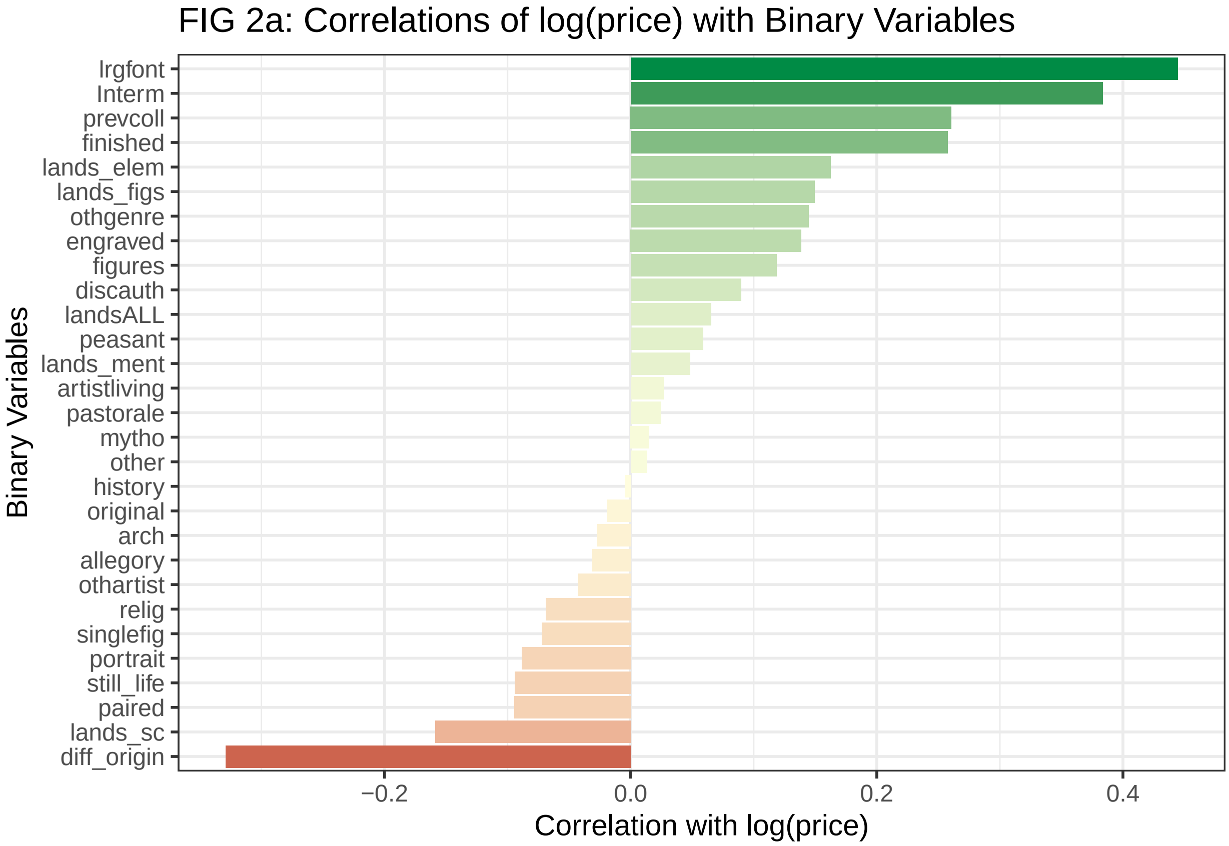


We see that the quantitative predictors plotted here are not really correlated with each other. Variable `position` only has a small negative and not necessarily linear correlation with the response variable. `year` has a stronger effect on `logprice` (.26) and we could consider a sort of overall "inflation" on paintings even though it is not linear nor monotonic. `Surface` is also positively correlated with `logprice` (.17). Note here that we used a log transformation on `Surface` in order to make its relationship with `logprice` linear. Finally, variable `nfigures` behave in a strange way. While having no figure does not seem to give any information on `logprice`, we can see that for paintings with at least one figure, more figures is correlated with higher price. It will be interresting to explore the interraction of `nfigures` with some of the binary predictor variables that we explore next.

Binary Variables

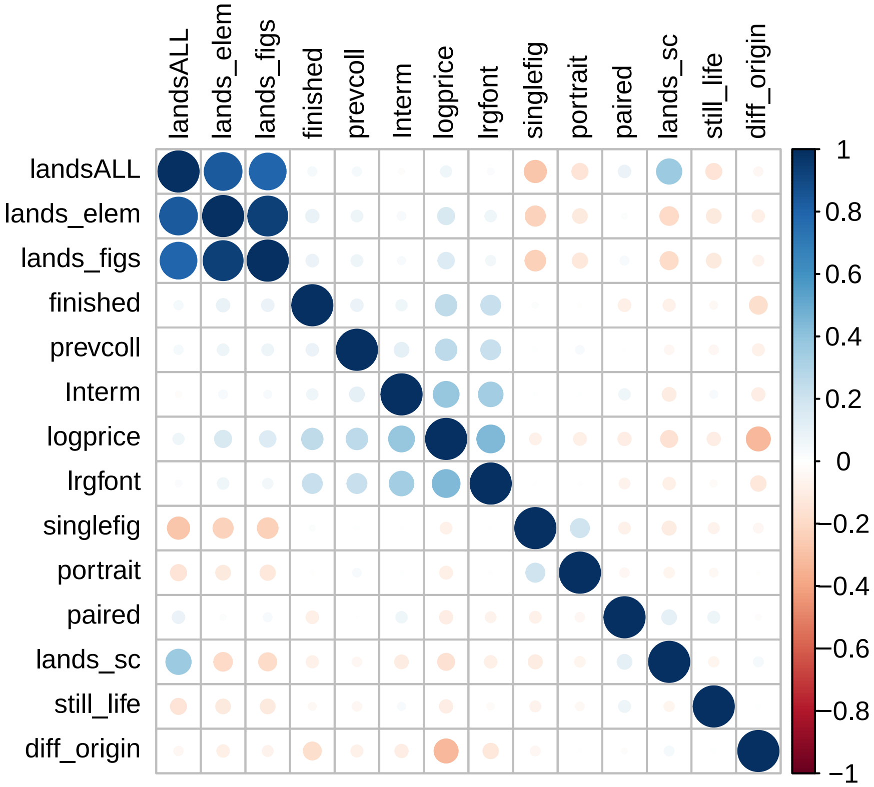
`diff_origin`, `artistliving`, `Interm`, `figures`,`engraved`, `original`, `prevcoll`, `othartist`, `paired`, `finished`, `lrgfont`, `relig`, `landsALL`, `lands_sc`, `lands_elem`, `lands_figs`, `lands_ment`, `arch`, `mytho`, `peasant`, `othgenre`, `singlefig`, `portrait`, `still_life`, `discauth`, `history`, `allegory`, `pastorale`, `other`

We present correlations between all the binary variables and log(price) (FIG 2a). Unsurprisingly, an additional paragraph in a larger font has a high correlation with price, suggesting that these paintings were the highlights in the different auctions. An intermediary also suggests a higher price, maybe because these individuals are involved mostly in high stakes sales. A mention of the previous owner and having a highly polished finishing are also factors that seem to drive up the price. Factors the drive the prices down are different origin of author and painting, if the content of the painting includes a "plain landscape", if the painting is just a "pairing" of another art work, and if the content is still life.



Once we have an idea of the top binary variable candidates to include in our model, it is important to take a look at their correlations within themselves and with other variables. Fig 2b is a correlation matrix of select binary variables. From this plot, it is clear that there is no need to include more than one variable among `landsALL`, `lands_elem` and `lands_figs` as they are all related in content, highly correlated and have similar effect on prices. `singlefig` is strongly negatively correlated with the above three, and negatively correlated with price, so we should also consider dropping it if we select one of the others. Regarding `lrgfont` and `Interm`, although both of them have a high correlation with prices, they are also correlated with each other, so we might consider choosing one of them.

FIG 2b: Correlation plot of log(price) and select binary variables

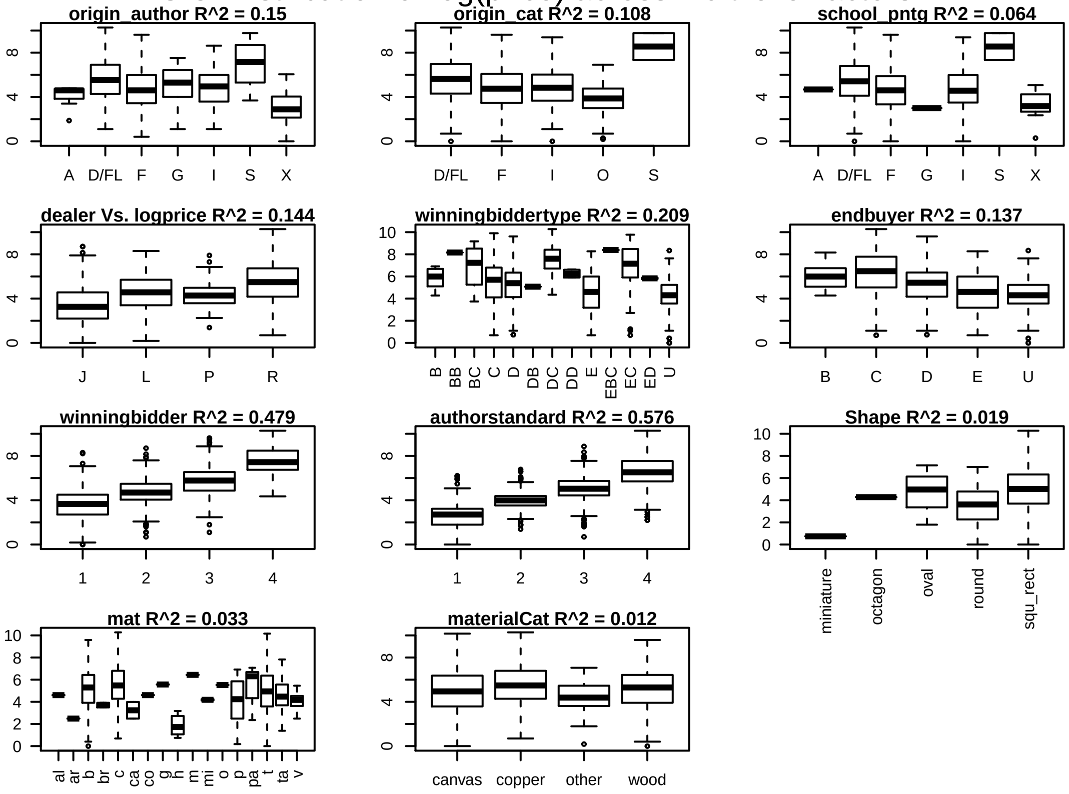


Multiple Level Factor Variable

`year`, `origin_author`, `origin_cat`, `school_pntg`, `Shape`, `material`, `mat`, `materialCat`, `dealer`, `winningbiddertype`, `endbuyer`, `type_intermed`

We choose here to include quartiles of `winningbidder` and `authorstandar` made in the data preprocessing based on the average buying/selling `logprice` of each winning bidder/ author. The scale goes from 1 least expensive to 4 most expensive. When investigating the material variable, we decide to exclude `material` as it has too many levels. As it has an important number of NA values, `type_intermed` is also ignored for now but will be considered later for interactions. (FIG 3)

FIG 3: Distribution of log(price) across multilevel factors



Using intuition, we could classify the first three plots regarding `origin_author`, `origin_cat` and `school_pntg` as information about the author. We expect these variables to be correlated with each other and therefore only using one of them would most likely give us enough information. We decide here to select `origin_author` as it has the highest r.squared (.15). The next three plots give us information about the sale of each painting. Looking at the `dealer` plot, we see that sale prices seem to be a little different across dealers. This might be explained by the kind of painting they each sale or by the kind of client they reach to. Now focusing on the `winningbiddertype` and `endbuyer` plots, we can deduce that these two variables inform on the buyer and so are probably highly correlated with each other. We can observe some differences accross buyers. We think that an explaination for these differences might be the intervention of intermediaries. Looking at both of our quartiles plots, we see that they provide us with a lot of information about the price of the painting depending on the buyer or the seller. Finally, the last three plots can be categorized as shape and material. These plots only explain an insignificant amount of the variance in `logprice`. However, the effect of a different shape as well as the interaction between the `Shape` and `Surface` might be worth looking at in our model building. Regarding material, the interaction with `Surface` could be of interest.

From our EDA, we are able to extract what we consider the 10 most important variables to predict `logprice`. Looking at the quantitative variables `year` and `Surface`, one can see that they are quite strongly correlated with our response variable (r = .26 and .17) and bring information about price evolution across years and accross the overall size of the painting. For the variables giving us information on the author, `diff_origin` has a relatively strong correlation (about .33) with `logprice`. Variables `origin_author`, `origin_cat` and `school_pntg` are correlated with each other and some of their imformation is already carried by `diff_origin`. However, the variable giving us the most information about prices accros authors is our quartiles variables `quartiles_authors` with an r-squared of .58. Therefore, we decide to choose `diff_origin` and `quartiles_authors`. When looking at the sale of the paintings, one can observe that `winningbiddertype` has an r-squared of .21 when regressed on `logprice`. However, we can assume that it will be strongly correlated to `Interm` and therefore would select `endbuyer` (r.sq = .14) instead. Nonetheless, our new variable `quartiles_winningBidder` seems to outperform all of these variables with an r-squarred of .48. Even if it migght be correlated with `quartiles_authors` or `dealer` we still believe it is an important variable. Variable Interm (presence of an intermediary) is also quite strongly correlated with the response (about .38). Despite its high correlation with other variables included in this list, we still believe that it provides important information. Finally, when investigating characteristics of paintings, variables lrgfont is the one with the highest correlation with the response (>.4) and despite its correlation with Interm, we consider it as an important predictor. Variables lands_sc, prevcoll, finished are also considered important in predicting logprice with correlations between .15 and .25 and a low correlation with the other variables selected.

Our 10 variables:
lands_sc, prevcoll, finished, lrgfont, quartiles_authors, quartiles_winningBidder, Interm, Surface, year, diff_origin

Preliminary Model

This discussion is based on how our final model from Part-I should have performed under the test dataset uploaded on December 12th 2019. In part A of the project, we ran stepwise selection on the 10 variables we indentified as important in the Part-I EDA (lands_sc, prevcoll, finished, lrgfont, origin_author, endbuyer, interm, Surface, year, diff_origin), along with 10 additional variables (dealer, Shape, material_Cat, lands_figs, peasant, engraved, portrait, still-life, discauth, artistliving) and a few interactions (log(Surface + 1):materialCat, dealer:origin_author, dealer:Interm, endbuyer:Interm, dealer:endbuyer and finished:prevcoll) that looked relevant in the data . The initial model included 77 coefficients, since many of our variables are categorical and were transformed into dummy variables in the regression. After stepwise selection with AIC as the selection criteria and two-directional search, we ended up with 36 coefficients or 19 variables (lands_figs was dropped) and one interaction (finished:prevcoll). We chose this model as our model for prediction.

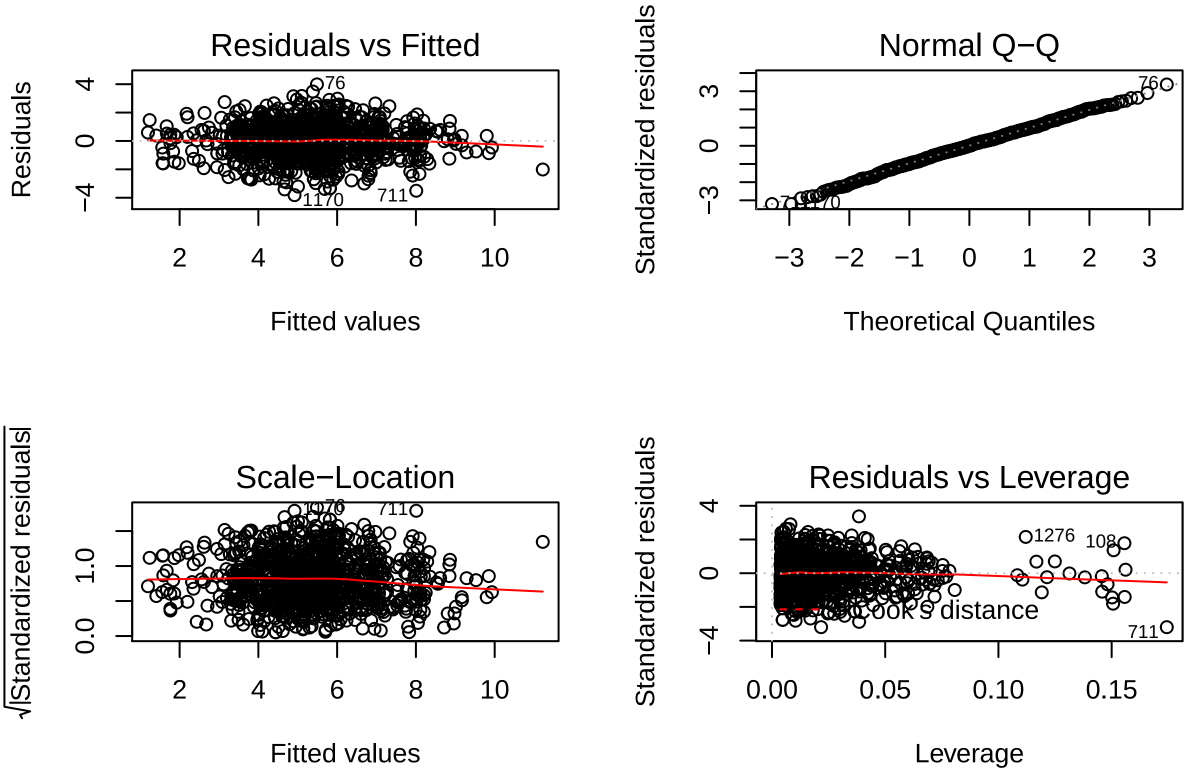
Looking back, we first see that our model performed quite well with a coverage of 95.3%, a bias of 234.07 and a RMSE of 1278.16. We could however have gone deeper in the EDA part by for example testing year as a factor variable or exploring variables such as authorstandard and winningbidder. Even though our bias and RMSE are quite low, better variable selection in the initial model would probably have helped us reduce them even more. To remedy this issue, we should have relied more on our EDA and limit the number of variables in our initial model. In order to investigate relevant interaction, we could have ran a stepwise regression on these selected variables with all two way interactions and pick the ones that seemed the most relevant. An other possibility was to use a different criteria for model selection. On the same initial model, running stepwise selection with BIC instead of AIC results in a smaller model with only 24 coefficients. Comparing the two options, BIC has a lower R^2 (0.6 compared to 0.614) and a higher residual se (1.19 compared to 1.17). Finally, We could have improved the model using cross validation.

Development of Final Model

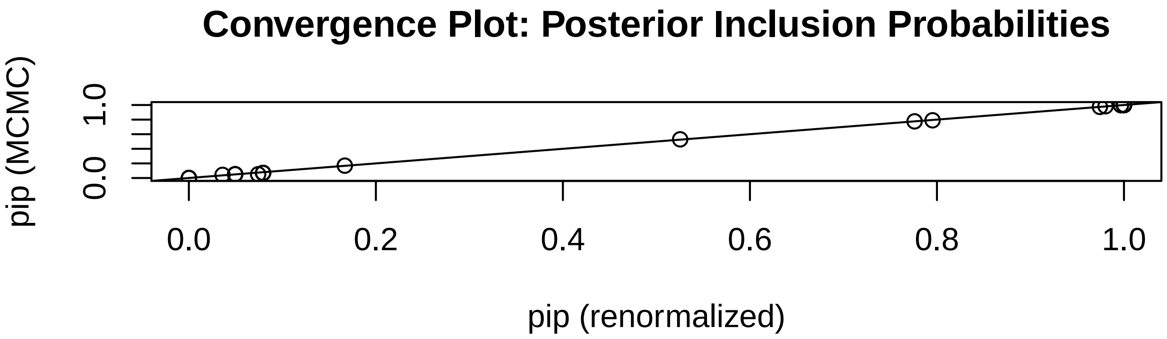
From what we learned from the first part, we decide to pick a more restricted set of variables for our initial model. Using k-fold cross validation with k=5, we try different selections including the variables present in our model from Part-I as well as year as factor and our two quartiles variables. We realize however, that despite a better RMSE in the painting_test dataset year as a factor seem to add instability in the model that we can observe through k-fold cross validation. Regarding the quartiles, they seem to reduce our out-of-sample performance consistently. We therefore end-up with a subset of 15 variables that we used in our model in Part-I (log(Surface), year, Interm, engraved, prevcoll, finished, lrgfont, lands_sc, portrait, still_life, discauth, artistliving, dealer, origin_author and materialCat). In order to investigate potential interactions, we run a backwards stepwise regression with AIC as selection criteria on our selected variables with all two way interactions. From this stepwise model, we pick the interactions that seem the most relevant (year:discauth, year:artistliving, Interm:discauth, Interm:dealer, prevcoll:finishedand discauth:dealer).

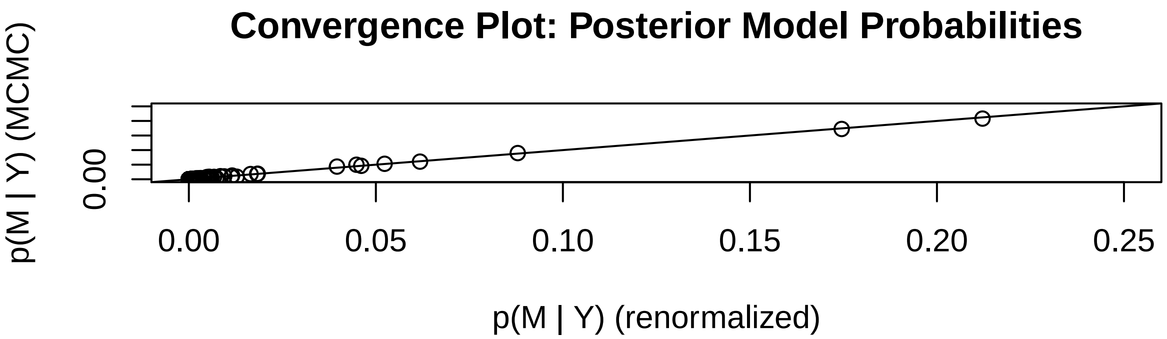
From this initial model, we choose to use BMA to produce our final model for this task. The advantages of using BMA are that it tends to perform better in terms of out-of-sample prediction compared to single-model techniques. It accounts for out-of-sample uncertainty and includes a version of selection by shrinking unnecessary coefficients towards zero. Markov Chain Monte-Carlo used in this analysis also avoid to enumerate every model and focuses on the one with higher probabilities.

Since we want to restrict the model to only the most important variables, we feel it is better to go with a more complex technique that will perform another kind of variable selection and reduce the tendency of linear models to overfit. In addition, BMA seems relevant in this case compared to tree-based techniques since the response is a continuous variable that should be monotonically correlated with the coefficients. Lasso and Ridge were also considered but we decided against them because estimating standard errors is more natural in a BMA setting and show higher out-of-sample RMSE. We use Highest Probability Model as our estimator as, it performed better (lowest test RMSE) than BMA and BPM in cross-validation as it will be shown in the evaluation of the model. Intervals are estimated from the posterior thanks to the `predict.bas` function using the HPM model. It calculates intervals the same way as a linear regression model



Above are the residual plots for the HPM model. They show that the model meet assumptions for linear regression (equal variance, normal residuals, etc.). We can then pursue with this model.

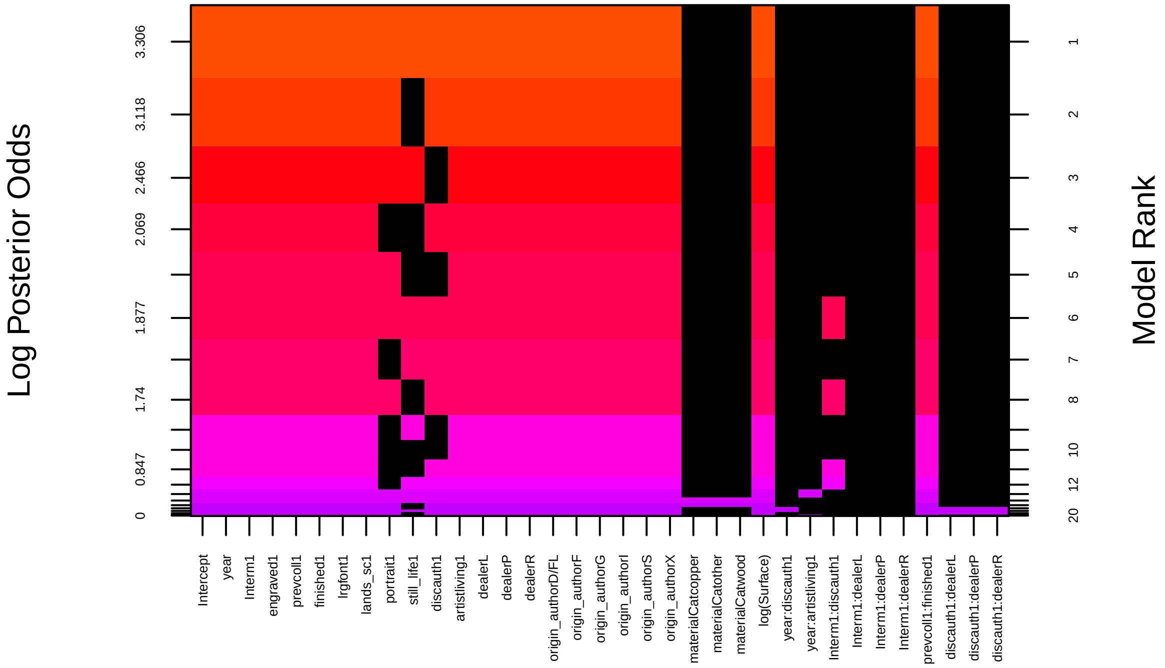




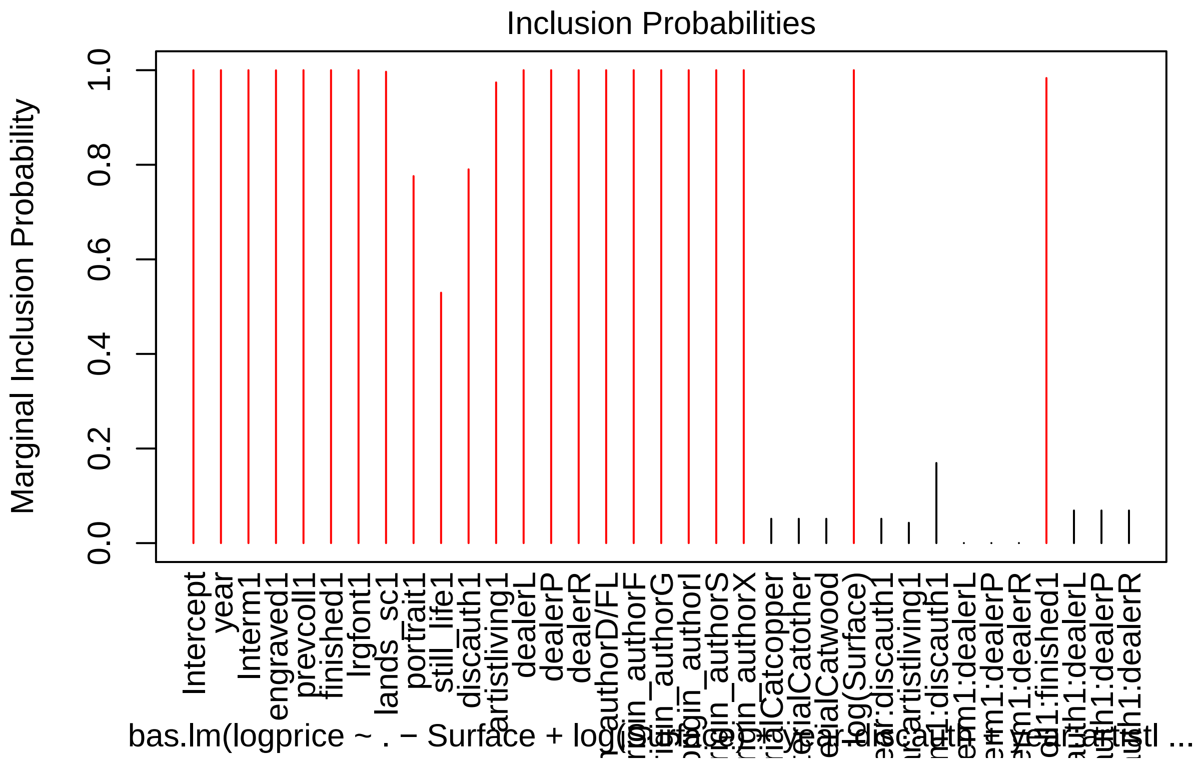
The diagnostics plots for the BMA shows that the models converged. The dots along the diagonal show that the estimates of the marginal inclusion probabilities from the re-normalized posterior odds agree with the estimates based on Monte Carlo frequencies. This means that enough MCMC samples were obtained.

	P(B != 0 | Y)	model 1	model 2	model 3	model 4	model 5
Intercept	1.000	1.000	1.000	1.000	1.000	1.000
year	1.000	1.000	1.000	1.000	1.000	1.000
Interm1	1.000	1.000	1.000	1.000	1.000	1.000
engraved1	1.000	1.000	1.000	1.000	1.000	1.000
prevcoll1	1.000	1.000	1.000	1.000	1.000	1.000
finished1	1.000	1.000	1.000	1.000	1.000	1.000
lrgfont1	1.000	1.000	1.000	1.000	1.000	1.000
lands_sc1	0.997	1.000	1.000	1.000	1.000	1.000
portrait1	0.776	1.000	1.000	1.000	0.000	1.000
still_life1	0.529	1.000	0.000	1.000	0.000	0.000
discauth1	0.790	1.000	1.000	0.000	1.000	0.000
artistliving1	0.974	1.000	1.000	1.000	1.000	1.000
dealerL	1.000	1.000	1.000	1.000	1.000	1.000
dealerP	1.000	1.000	1.000	1.000	1.000	1.000
dealerR	1.000	1.000	1.000	1.000	1.000	1.000
origin_authorD/FL	1.000	1.000	1.000	1.000	1.000	1.000
origin_authorF	1.000	1.000	1.000	1.000	1.000	1.000
origin_authorG	1.000	1.000	1.000	1.000	1.000	1.000
origin_authorI	1.000	1.000	1.000	1.000	1.000	1.000
origin_authorS	1.000	1.000	1.000	1.000	1.000	1.000
origin_authorX	1.000	1.000	1.000	1.000	1.000	1.000
materialCatcopper	0.051	0.000	0.000	0.000	0.000	0.000
materialCatother	0.051	0.000	0.000	0.000	0.000	0.000
materialCatwood	0.051	0.000	0.000	0.000	0.000	0.000

	P(B != 0 | Y)	model 1	model 2	model 3	model 4	model 5
log(Surface)	1.000	1.000	1.000	1.000	1.000	1.000
year:discauth1	0.051	0.000	0.000	0.000	0.000	0.000
year:artistliving1	0.043	0.000	0.000	0.000	0.000	0.000
Interm1:discauth1	0.169	0.000	0.000	0.000	0.000	0.000
Interm1:dealerL	0.000	0.000	0.000	0.000	0.000	0.000
Interm1:dealerP	0.000	0.000	0.000	0.000	0.000	0.000
Interm1:dealerR	0.000	0.000	0.000	0.000	0.000	0.000
prevcoll1:finished1	0.983	1.000	1.000	1.000	1.000	1.000
discauth1:dealerL	0.069	0.000	0.000	0.000	0.000	0.000
discauth1:dealerP	0.069	0.000	0.000	0.000	0.000	0.000
discauth1:dealerR	0.069	0.000	0.000	0.000	0.000	0.000
BF	NA	1.000	0.823	0.414	0.291	0.247
PostProbs	NA	0.208	0.172	0.090	0.060	0.053
R2	NA	0.605	0.603	0.602	0.600	0.600
dim	NA	23.000	22.000	22.000	21.000	21.000
logmarg	NA	547.983	547.788	547.102	546.749	546.583



Both the table and the model space show that the BMA procedure produces clear decisions on what should and what shouldn't be in the model. Out of our candidate variables, several were dropped from all top models. Others got probability of one to be kept in the model.



The Inclusion Probabilities graph is another visualization that helps us understand what the BMA procedure considers should and what shouldn't be kept in the final model.

Assessment of Final Model

Model Evaluation

Table 2: Posterior summaries of coefficients in final model

	post.mean	post.sd	post.P.B....0.
Intercept	4.9919	0.0326	1.0000
year	0.1245	0.0075	1.0000
Interm1	1.0644	0.1194	1.0000
engraved1	0.7138	0.1501	0.9997
prevcoll1	1.1456	0.1752	1.0000
finished1	0.9838	0.0978	1.0000
lrgfont1	1.0172	0.1233	1.0000
lands_sc1	-0.5589	0.1261	0.9967
portrait1	-0.5487	0.1722	0.7760
still_life1	-0.5005	0.1827	0.5295
discauth1	0.4249	0.1427	0.7904
artistliving1	0.4085	0.1072	0.9741
dealerL	1.0732	0.1294	1.0000
dealerP	0.2104	0.1644	1.0000
dealerR	1.9297	0.1085	1.0000
origin_authorD/FL	0.1677	0.4583	1.0000
origin_authorF	-0.5376	0.4597	1.0000
origin_authorG	-0.2643	0.5148	1.0000

	post.mean	post.sd	post.P.B....0.
origin_authorI	-0.6559	0.4665	1.0000
origin_authorS	-0.5656	0.6010	1.0000
origin_authorX	-1.3206	0.4699	1.0000
materialCatcopper	0.0000	0.0000	0.0514
materialCatother	0.0000	0.0000	0.0514
materialCatwood	0.0000	0.0000	0.0514
log(Surface)	0.3449	0.0280	1.0000
year:discauth1	0.0000	0.0000	0.0514
year:artistliving1	0.0000	0.0000	0.0430
Interm1:discauth1	0.0000	0.0000	0.1694
Interm1:dealerL	0.0000	0.0000	0.0001
Interm1:dealerP	0.0000	0.0000	0.0001
Interm1:dealerR	0.0000	0.0000	0.0001
prevcoll1:finished1	-1.2878	0.3231	0.9834
discauth1:dealerL	0.0000	0.0000	0.0690
discauth1:dealerP	0.0000	0.0000	0.0690
discauth1:dealerR	0.0000	0.0000	0.0690

Table 3: Performance of top 5 BMA models

	model 1	model 2	model 3	model 4	model 5
BF	1.0000	0.8226	0.4144	0.2913	0.2467
PostProbs	0.2080	0.1724	0.0898	0.0604	0.0530
R2	0.6051	0.6029	0.6025	0.6001	0.6000
dim	23.0000	22.0000	22.0000	21.0000	21.0000
logmarg	547.9829	547.7875	547.1020	546.7493	546.5831

From the posterior summaries above, we can see that most of the included coefficients have high marginal inclusion probabilities: only `materialCat` is below 0.5. Contrarily, the interaction terms, which were chosen based on an AIC stepwise selection, are mostly not significant: only `prevcoll:finished` is above 0.5. Based on the summary table, however, BMA was mostly able to find well-fitted models. The best model had an R^2 of 0.6051 and 23 coefficients, similar to our original linear regression model from Part I. Since for the sake of efficiency, we used MCMC to search the model space, there is no guarantee that we found the absolute best model, but of those searched, this “best” model has by far the highest posterior probability (0.2173).

Model Testing

To verify whether our BMA performs well both in-sample and out-of-sample, we used 5-fold cross-validation on the training set to compute the average RMSE and coverage across all folds. Below are the results for the BMA, BPM, and HPM estimators:

Table 4: k-fold CV for final model (BMA estimator)

	null_train	train_rmse	train_covg	null_test	test_rmse	test_covg
	2133.484	1441.398	0.947	2627.819	1915.932	0.944
	2154.045	1529.296	0.956	2557.785	1713.920	0.934
	2344.014	1582.288	0.947	1769.118	1311.500	0.959
	2182.964	1574.167	0.955	2457.883	1560.420	0.915
	2377.717	1605.646	0.955	1576.303	1541.233	0.967

	null_train	train_rmse	train_covg	null_test	test_rmse	test_covg
average	2238.445	1546.559	0.952	2197.781	1608.601	0.944

Table 5: k-fold CV for final model (HPM estimator)

	null_train	train_rmse	train_covg	null_test	test_rmse	test_covg
	2133.484	1421.483	0.946	2627.819	1948.392	0.933
	2154.045	1519.791	0.952	2557.785	1697.960	0.930
	2344.014	1578.597	0.948	1769.118	1310.456	0.959
	2182.964	1569.932	0.956	2457.883	1510.672	0.923
	2377.717	1612.914	0.951	1576.303	1277.595	0.967
average	2238.445	1540.543	0.951	2197.781	1549.015	0.942

Table 6: k-fold CV for final model (BPM estimator)

	null_train	train_rmse	train_covg	null_test	test_rmse	test_covg
	2133.484	1421.483	0.946	2627.819	1948.392	0.933
	2154.045	1519.791	0.952	2557.785	1697.960	0.930
	2344.014	1578.597	0.948	1769.118	1310.456	0.959
	2182.964	1569.932	0.956	2457.883	1510.672	0.923
	2377.717	1612.914	0.951	1576.303	1277.595	0.967
average	2238.445	1540.543	0.951	2197.781	1549.015	0.942

All 3 estimators seem to give similar results. This should be expected, as one model has a very high posterior probability, making BMA close to selection of the highest-posterior model. Based on these results, we decide to just use the HPM to make predictions, as it seems to perform the best of the three estimators on out-of-sample observations and as mentioned, it is naturally very close to the BMA estimates.

The highest-posterior model performs consistently well across all folds in both RMSE and coverage. It always does better than the null model and with the exception of the final fold, performs as well or better on the test set than on the training set. Of the different BMA models we fitted, we found that this final model has relatively low variance in test RMSE and on average, has the lowest test RMSE. Other models (e.g., one which used the same coefficients but treated `year` as a factor) performs better on the test set leaderboard (1232 RMSE instead of the current 1282 RMSE) but due to high variance, is much worse in cross-validation. Since we do not know how our performance on the validation set, we decide to choose the more consistent and stable model, despite the slightly lower score on the test set.

Top 10 Paintings

Table 7: Top 10 valued paintings in validation set

predict_price	author	subject
17765.73	Karel du Jardin	Marchand dorvitans en habit de Scaramouche scne de thtre specta
13639.27	Rembrandt Van Rhyn	Arquebusiers Ronde de nuit
13572.85	Pierre Paul Rubens	Adoration des Rois
13055.32	Joseph Vernet	Paysage une tempte PENDANTS
9358.79	Jacques Jordaens	LE ROI BOIT
8948.86	Philippe Wouwermans	une chasse au cerf

predict_price	author	subject
7450.95	Nicolas Berghem	Vue du chateau de Bentheim figures animaux
6428.04	Vander Heyden et Adrien Vanden Veld	Une des portes de Cologne figures
6419.33	Salvator Rosa	Paysage avec Tobie lange
6059.89	Nicolas Berghem	Figures de differentes nations homme joue de la guitare femme ba

Our model selects the above works as the top 10 most expensive paintings in the validation set. The second and third artists on this list are Rembrandt and Peter Paul Rubens, respectively, perhaps the two most well-regarded artists of the 17th century Dutch Golden Age. In particular, the second painting is "The Night Watch," Rembrandt's most famous painting and one of the most famous of that entire century. Jardin is less well-known nowadays, but since these paintings were bought in the late 18th century, it is possible his reputation was better at the time. Regardless, these results seem to make sense, suggesting that our model is giving reasonable results on out-of-sample predictions (at least at the high end of the price spectrum).

Conclusion

It seems like the factors that drove painting prices in 18th century Paris are mostly related to the sale and the dealer, the attributes of the painter and the year the sale took place. Only a few elements that relate to the actual content of the painting remained in our final model. The circumstances of the sale, therefore, have a higher prediction power than the subject that shows up in the painting. This outcome is not very surprising, and one would postulate that this is not a feature unique to pre-industrial France, but a persisting feature of the art world to this day.

Our modelling efforts included multiple trial and error on variable selection and modelling technique. We learned from this experiment that variables that seem to perform very well in the training data can result in overfitting the model when applied to out-of-sample data. For example, we found out that the year variable performs well in the test data when it is viewed as a categorical variable. But when performing k-fold cross-validation, it seemed to bring instability to our model with high variance in RMSE. When we included it as a numeric value (therefore considering a linear relationship with log price) the out-of-sample results got much better results. In addition, once the basic model is decided, using more advanced techniques such as BMA does not necessarily result in better predictions.

If we had more time, we would consider other modelling techniques such as BART, and also experiment more with other variables. For example, possibly applying vectorization of the content strings and identifying imporant attributes. Also, we would experiment with dimensionality reduction of the dummy variables before including them in the models, to see if more information could be drawn from them. Finally, we would have executed multiple imputation with `mice`.

Final Predictions