Statistical analysis of artwork sold at auction houses

Alessio Arcudi, Alessandro Mazzoni, Alessandro Padella

University of Padua
Second cycle degree in Data science

Academic year 2019-2020

Index

1. Obtaining data
2. Clean and filter data
3. Explore data
4. Model data
   - Predictor model
   - Classification model
5. Interpreting data
   - Comments on accuracy
6. Interpreting data
   - Comments on accuracy
7. Technical appendix

Introduction

This project is realized in a dataframe composed of featured related to lot of real artworks.
We downloaded all the informations contained in it.
In a first part we'll analyze some relation between features, in particular price related to other columns of the dataframe.
In the second part we'll build different models to predict and classify the prices of the artworks.

Downloading data with python

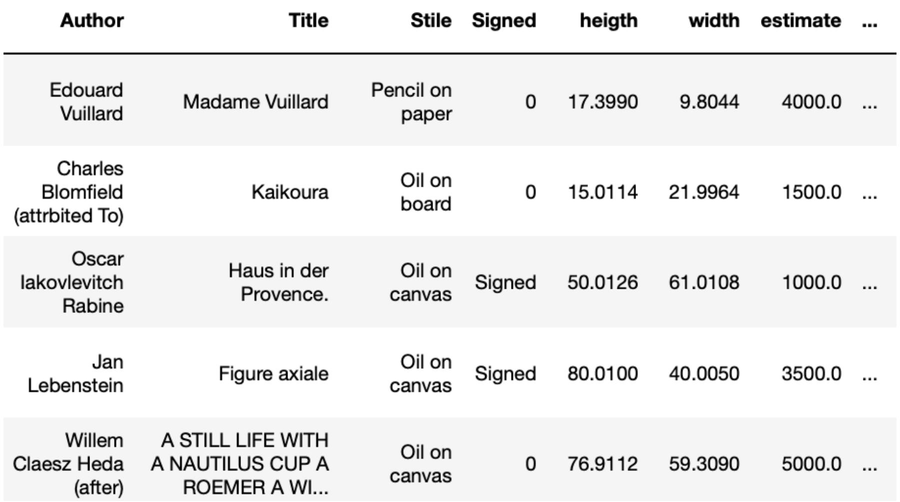
We used the python packets *request*[1] and *bs4*[2] *(Beautiful Soup 4)* in order to download and organize the data in a *.csv* format.
The first part of the algorithm is a request algorithm that establish a connection with the site server, instead the second part take the HTML obtained and look for data, moving throughout the HTML tags finding the one that describe the artwork.

[1] *The requests library is the de facto standard for making HTTP requests in Python. It abstracts the complexities of making requests behind a beautiful, simple API.*

[2] *Beautiful Soup is a Python library for pulling data out of HTML. It provides idiomatic ways of searching and modifying the HTML parse tree.*

At the end of the part the result provided was a raw dataset with 14 features such as:

- artwork title
- Author
- Price in the auction currency
- dimensions
- Currency
- *etc ...*



Author	Title	Stile	Signed	heigth	width	estimate	...
Edouard Vuillard	Madame Vuillard	Pencil on paper	0	17.3990	9.8044	4000.0	...
Charles Blomfield (attrbited To)	Kaikoura	Oil on board	0	15.0114	21.9964	1500.0	...
Oscar Iakovlevitch Rabine	Haus in der Provence.	Oil on canvas	Signed	50.0126	61.0108	1000.0	...
Jan Lebenstein	Figure axiale	Oil on canvas	Signed	80.0100	40.0050	3500.0	...
Willem Claesz Heda (after)	A STILL LIFE WITH A NAUTILUS CUP A ROEMER A WI...	Oil on canvas	0	76.9112	59.3090	5000.0	...

Finally we deleted immediately all the raws that contain *NaN* values and all the columns that contain sterile information, such as the Lot number or the identification number.
Moreover we transformed with python all the different prices and estimates in USD[1] in order to be comparable.
At the end of the part of obtaining data we have a first row dataset of 5000 artworks (`art.csv`) that needed to be cleaned and refined in order to pass it to a computational part.

[1] from now on we will refer to the prices of the different artwork as price in the USD currency

Categorical Data

House action : this feature is referred to the auction house in which an artwork was sold, we maintained it as it was originally.

currency : This is the currency in which the artwork was sold, pay attention to the fact that not every-time it is the currency used in the place in which it was sold and it's not true that it's directly linked to the auction house.

Signed : this is a dihcotomic feature that is 0 if the artwork it's not Signed and 1 if it is.

date : this is an ordinal categorical data which describe when the artwork was sold, we took the most old date $d_o = 0$ $(10/10/2001)$ and set to 0, then the other dates transformed $d_i - d_o$ counted in days.

- **Style** : this feature is referred to the style of the artworks, we noticed that the original style feature has as first word of the description a keyword that describe it generically and the other words were a further insight of style of the object (*Oil on canvas -> Oil, Watercolor over pencil-> Watercolor*).
- **Place** : this feature is referred to the nation in which it was sold, originally this feature had in the same string the city and the nation, we decided to generalise it better taking only the last word of the string [1]

[1] before doing so we checked if there was problem with states with two word such as *United Kingdom* but referring to it with only *Kingdom* didn't create ambiguity with other nations.

With regard to the variables **Style**, **Place** and **House action** in order to create general categories without connect directly a category to one piece of art we decided to set a new category called 'Other' for these variables that contain all the feature that appear less than 19 times.

Numerical data

height and width : height and width of the artwork in cm.

estimate : estimate price of the artwork done by the auction house.

price : price at which the artwork was sold.

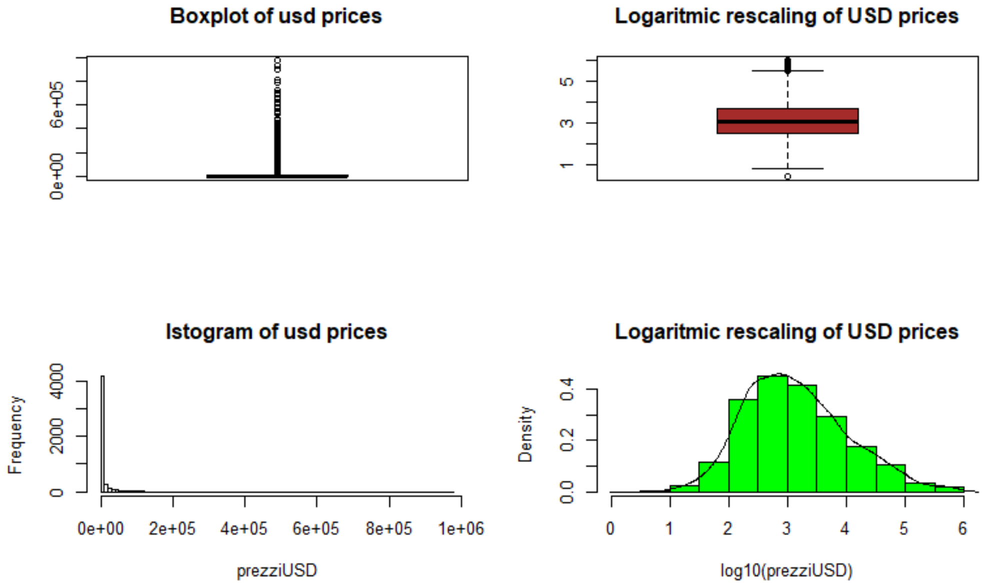
This numerical data had been refined previously in python thanks to python packet that converted inche to centimetre and prices from one currency to another in the specific date in which they were sold.

Quantitative analysis: Prices

Since our variables are both quantitative and categorical, we used two different methods to treat them.

- **Price**: The first approach, after to have done summaries,has been plot the boxplot and histograms but they're not centered,

it's due to the fact that there are values too far from the mean, so we decided to apply a log-trasformation to better visualize the distribution.

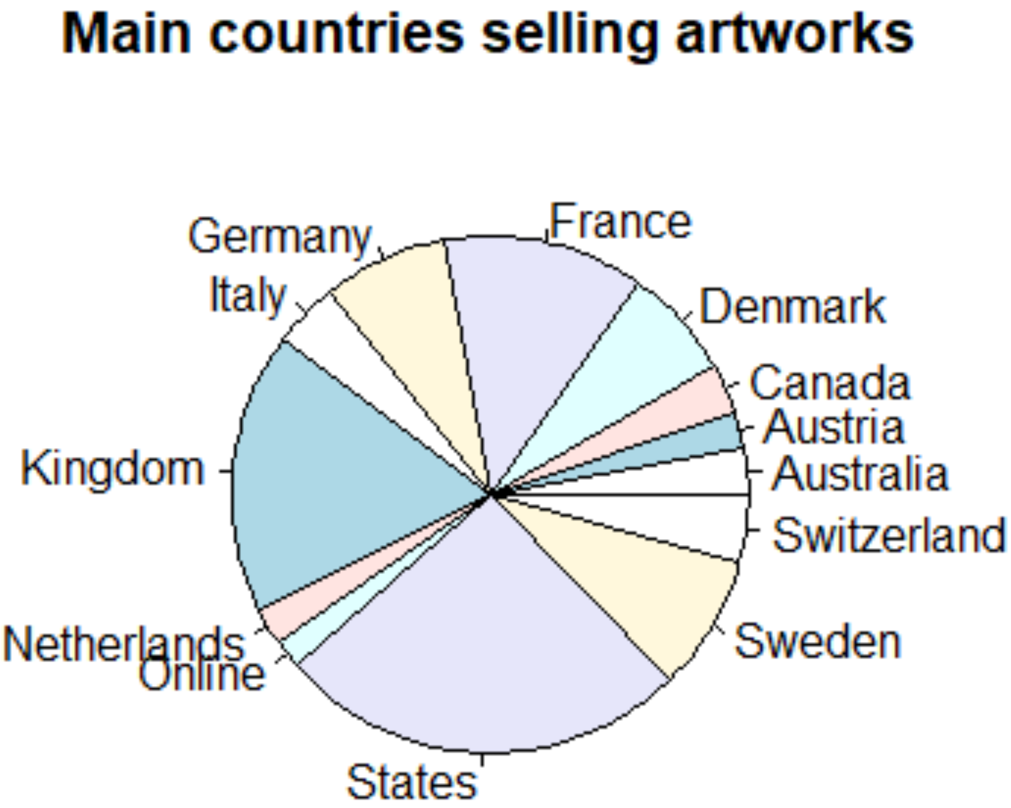


Quantitative analysis: Height and width

- **Height and width** The analysis of height and width was close to the analysis of prices, we had to apply the logarithm to rescale data.
  A deeper analysis with a t.test showed that the artworks width is usually higher than artworks height.
  We also tried to check if the height and width datas come from a normal distribution using the Shapiro-Francia test and we could not accept the hypothesis
- **Estimate** We compared the values on the Estimate column with the values in the price column and we evaluated the mean squared error of the estimates.

Qualitative analysis: Stile, place

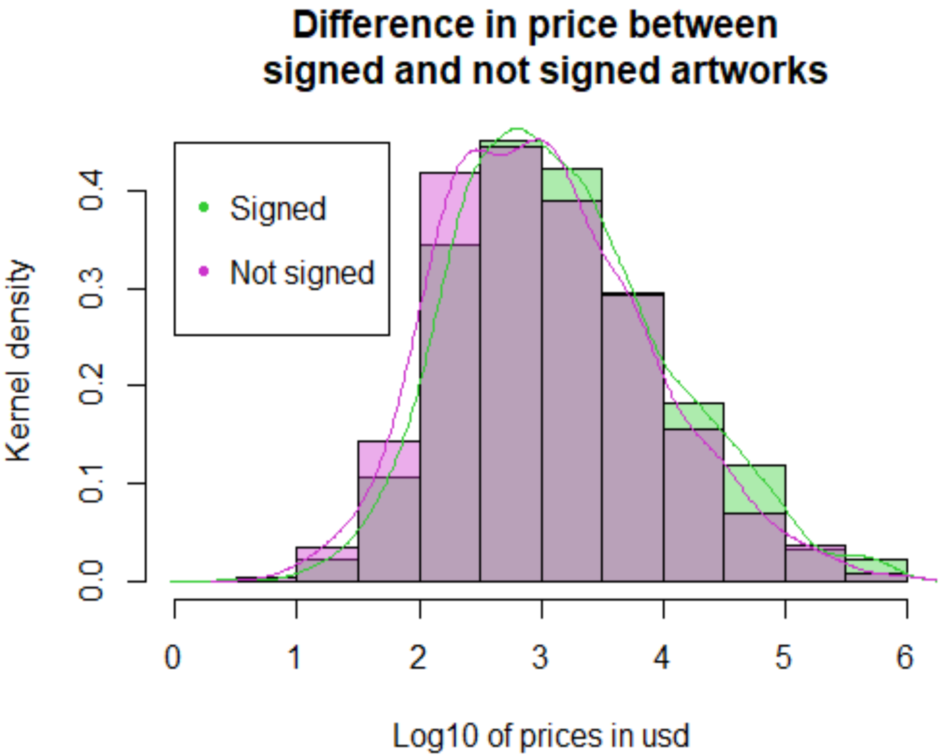
- **Stile** The datas were pre-processed so we represented the preprocessed datas in a pie chart
- **Place** The datas were pre-processed so we represented the preprocessed datas in a pie chart and we tried to find the countries in which the selling-prices are high using creating a boxplot that relates the most popular countries and the selling-prices.
- **Auction houses** The datas were pre-processed so we represented in a pie chart the distribution of auction houses


Main countries selling artworks
France
Denmark
Canada
Austria
Australia
Switzerland
Sweden
States
Online
Netherlands
Kingdom
Italy
Germany

Singature variable

- **Signature** is a dichotomic variable, the 77% of the artworks are signed.

We also noticed that the signature influence a lot the price of an artwork and we represented in an histogram with the related density distribution.



Goal

The goal of this work is to build a model able to *predict* artwork prices.

The variable prices is a numerycal variable, so the first approach chosen to solve this problem is with a **linear regressor**.

We'll begin with a model with few variables, increasing (while it's possible and the resulting model will be satisfying) them a little at a time.

1^{st} model: description

The first model is:

$$Y_i = \beta_0 + \beta_1 x_{i,h} + \beta_2 x_{i,w} + \beta_3 x_{i,s} + \sum_{j \in S} \beta_j x_{i,j}$$

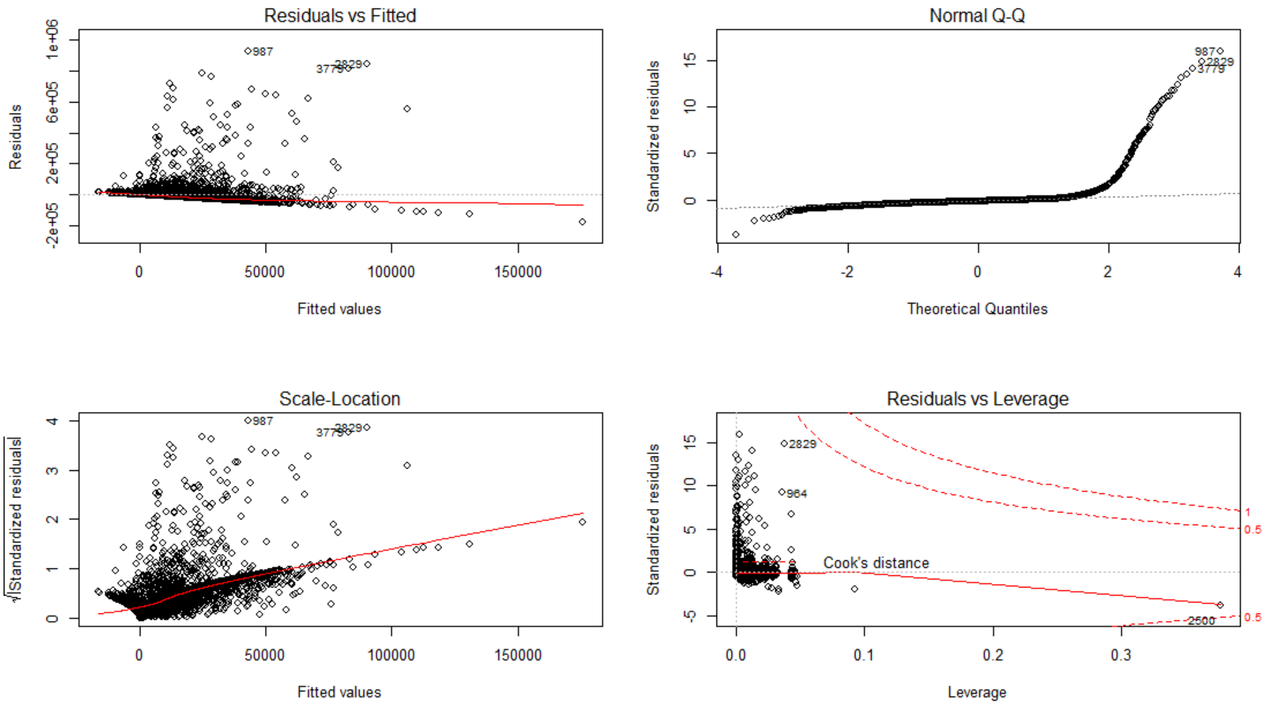
where:

- $x_{i,h}$ is the variable that represents the height of the i^{th} artwork;
- $x_{i,w}$ is the variable that represents the width of the artwork;
- $x_{i,s}$ is the dummy variable that represents if the i^{th} artwork is signed $(x_{i,s} = 1)$ or not $(x_{i,s} = 0)$;
- $x_{i,j}$ is the variable that represents if the i^{th} artwork belongs to the style j.

Remember that $|S|$ (the nr.of styles) is 30, so this model has 33 coefficients β_k.

1^{st} model: diagnostics

The value of $R^2 = 0.05319$ and of the adjusted $\hat{R}^2 = 0.0471$



2^{nd} model: description

The second model is:

$$\log(Y_i) = \beta_0 + \beta_1 \log(x_{i,h}) + \beta_2 \log(x_{i,w}) + \beta_3 x_{i,s} + \sum_{j \in S} \beta_j x_{i,j}$$

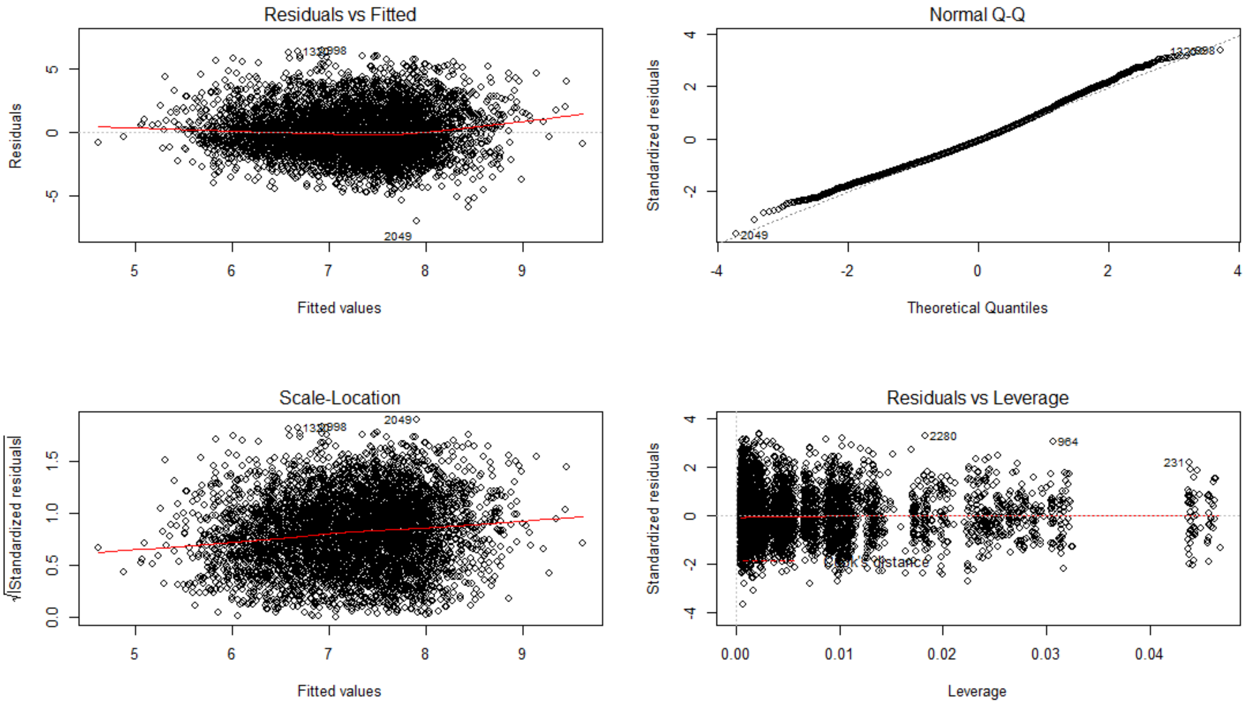
where:

- $x_{i,h}$ is the variable that represents the height of the i^{th} artwork;
- $x_{i,w}$ is the variable that represents the width of the artwork;
- $x_{i,s}$ is the dummy variable that represents if the i^{th} artwork is signed $(x_{i,s} = 1)$ or not $(x_{i,s} = 0)$;
- $x_{i,j}$ is the variable that represents if the i^{th} artwork belongs to the style j.

As above, this normalized model has 33 coefficients β_k.

2^{nd} model: diagnostics

The value of R^2 is 0.1132 and the of adjusted $\hat{R}^2$ one is 0.1075: there is a little improvement in the model.



3^{rd} model: description

The third model is:

$$\log Y_i = \beta_0 + \beta_1 \log x_{i,h} + \beta_2 \log x_{i,w} + \beta_3 \log x_{i,s} + \tag{1}$$

$$+ \beta_4 x_{i,d} + \sum_{j \in S} \beta_j x_{i,j} + \sum_{p \in P} \beta_p x_{i,p} \tag{2}$$

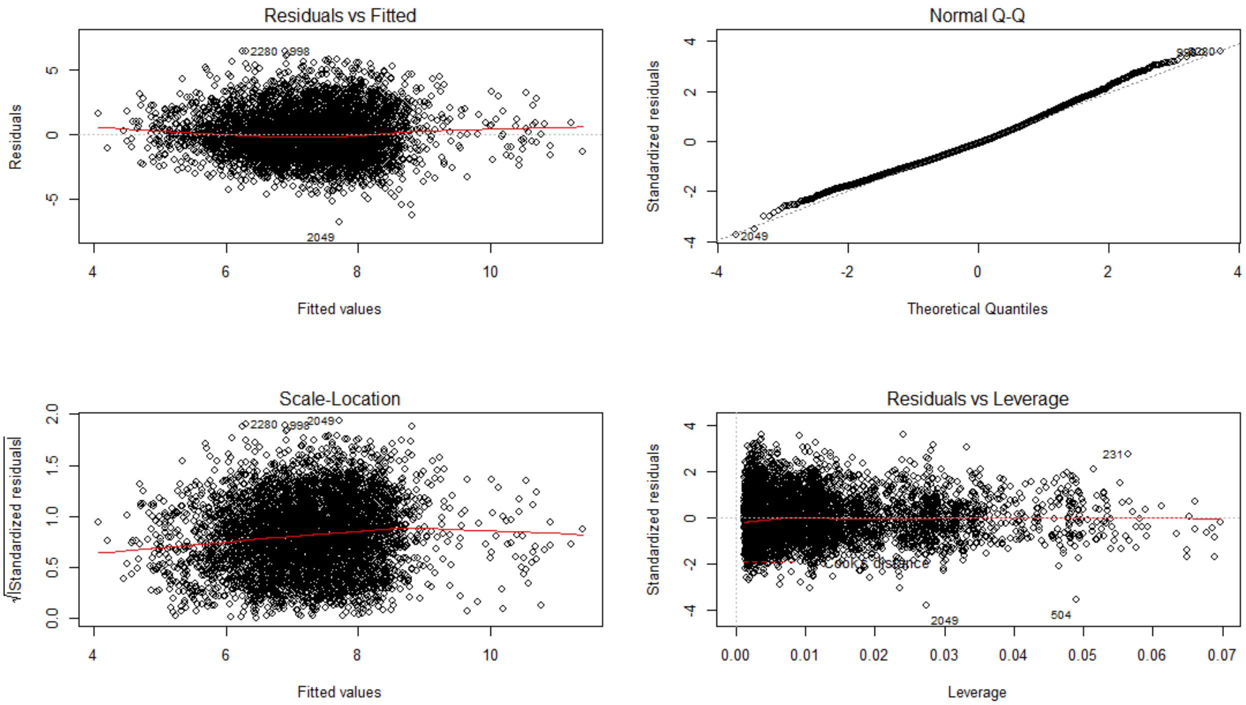
where:

- $x_{i,h}, x_{i,w}, x_{i,s}$ and $x_{i,j}$ are the same variables presented in the previous slides;
- $x_{i,d}$ is the variable that represents date of the artwork;
- $x_{i,p}$ is the variable that represents if the i^{th} artwork has been sold in the place j.

The fact that $|S| = 30$ and $|P| = 27$ implies that this models has 60 coefficients β_k.

3^{rd} model: diagnostics

The value of $R^2 = 0.194$ and of the adjusted $\hat{R}^2 = 0.1844$



4^{th} model: description

Here the fourth model:

$$\begin{aligned} \log Y_i = \beta_0 + \beta_1 \log x_{i,h} + \beta_2 \log x_{i,w} + \beta_3 \log x_{i,s} + \beta_4 x_{i,d} + & \quad (3) \\ + \sum_{j \in S} \beta_j x_{i,j} + \sum_{p \in P} \beta_p x_{i,p} + \sum_{r \in H} \beta_h x_{i,h} + \sum_{c \in C} \beta_c x_{i,c} & \quad (4) \end{aligned}$$

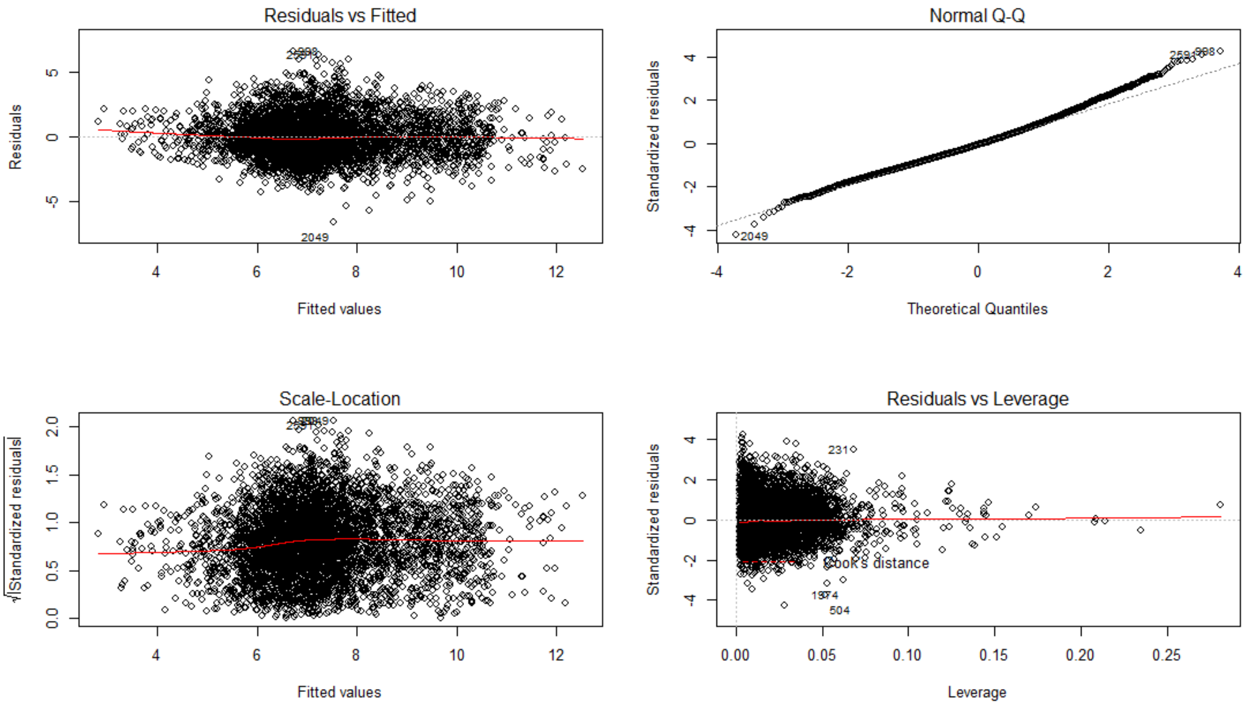
The only new variables presented here are:

- $x_{i,r}$ is the variable that represents if the i^{th} was sold in the auction house r;
- $x_{i,c}$ is the variable that represents if the i^{th} was sold with currency c.

$|H| = 49$, $|C| = 16$ so we have 63 coefficients in more.
The final model has 123 coefficients β_k.

4^{st} model: diagnostics

Here $R^2 = 0.4102$ and the adjusted $\hat{R}^2 = 0.3958$



LR: model analysis

We splitted the dataset in this way:

- 90% of the dataset (that is the first 4500 samples) became the **training set**;
- 10% of the dataset (the last 500 samples) became the **test set**;

We made the regression (chosing the 4^th model presented above) over the **training set** and then we predicted the log-price values for the samples in **test set** according to the coefficients computed in the learning phase. Our

goal was to compare the vector of predicted values y with the vector of estimates e.

LR: model analysis

We saw that:

$$\frac{1}{500}\sum i = 1^{500}|\log(pi) - \log(e_i)| = 0.4597632$$

$$\frac{1}{500}\sum i = 1^{500}|\log(pi) - y_i| = 1.22676$$

where y_i is the predicted value, e_i the estimate and p_i the actual price for each artwork in the test set.

The model seems to perform worse than the estimates by the action houses. We made a paired t-test among the two population, and it gave a big evidence ($p - value < 2 \times 10^{-16}$) to reject the null hypotesis that the first mean is bigger than the second one.

Classification model: Goal

We extended the features with a column "fascia" that represent the price range of the artworks.
The goal of this part of the project is creating a model that is able to distinguish if an item is an expensive masterpiece or not.
To approach this classification problem we chose to use Linear Discriminant Analysis[1], a multi-classifier which fitted well for this task.

[1] is a method used in statistics, pattern recognition, and machine learning to find a linear combination of features that characterizes two or more classes of objects. The resulting combination may be used for dimensionality reduction before later classification.

- We begun with a classification over 5 price ranges but the model was very fragmented and didn't lead to good results.
- Then we tried splitting the prices in 3 ranges: 0-1000$, 1000-50000$ and prices over 50000$. The resulting confusion matrix was a better classification.

Confusion Matrix

	1	2	3
1	0.811	0.427	0.241
2	0.177	0.452	0.290
3	0.012	0.121	0.469

Then we noticed that the artworks were not equally distributed between the 3 classes, but the third class was only the 5% of the all dataframe, instead the other 2 were above 40%.

Secondly, we decided to overcome this issue by applying the `SMOTE`[1] function. The new generated dataset is now equilibrated with the same number of sample for each range. The LDA applied to this dataset:

confusion matrix

	1	**2**	**3**
1	0.716	0.302	0.087
2	0.256	0.472	0.186
3	0.028	0.226	0.727

LDA reduced parameters

[1] iThis function handles unbalanced classification problems using the SMOTE method. Namely, it can generate a new "SMOTEd" data set that addresses the class unbalance problem. Alternatively, it can also run a classification algorithm on this new data set and return the resulting model.

We finally tried to divide the model in a train set and validation set in order to evaluate if it maintains its properties. We took a train set of 5000 random elements and a test set of the remaining 830, the evaluation of the test set led to:

confusion matrix

	1	**2**	**3**
1	0.678	0.318	0.077
2	0.297	0.432	0.163
3	0.0247	0.249	0.760

LDA reduced parameters

We also evaluated it on a test set of the original dataset, that has less than 5% of elements in the third class.

confusion matrix

	1	2	3
1	0.735	0.318	0.103
2	0.234	0.448	0.231
3	0.031	0.234	0.667

Comments of accuracy

- For the predictor we reached an $R-$squared index of 0.4 and for the classifier an average accuracy mean of 65 $\%$.
- In literature there are other attempts to predict and classify the value of artworks, also using the painting features (color, patterns, lines, style ecc) and with other approaches but everyone agreed has proven that is still difficult to achieve good rates of precision when algorithm has to deal with human sensation.
- Other researches suggest that extrinsic features like artists' reputation and the number of bidders are more important than features unique to the artwork.

Possible uses

- So even if it may seem not a good result it can help the buyers/sellers to state a certain point to start and how to orient their purchase/vending strategy.
- This analysis and also suggest to buyers/sellers not to consider just features like style and artist but also consider the currency and the auction house when they have to choice how to split their budget

Comments of accuracy

- For the predictor we reached an $R-$squared index of 0.4 and for the classifier an average accuracy mean of 65 $\%$.
- In literature there are other attempts to predict and classify the value of artworks, also using the painting features (color, patterns, lines, style ecc) and with other approaches but everyone agreed that is still difficult to achieve good rates of precision when algorithm has to deal with human sensations.
- Other researches suggest that extrinsic features like artists' reputation and the number of bidders are more important than features unique to the artwork.

Possible uses

- So even if it may seem not a good result it can help the buyers/sellers to state a certain point to start and how to orient their purchase/vending strategy.
- This analysis and also suggest to buyers/sellers not to consider just features like style and artist but also consider the currency and the auction house when they have to choice how to split their budget

We applied two different approach:

- A classical regression approach, that has been done considering first of all datas. After some attempts we found that, in order to diversify, to execute a logarithmic rescaling is the better option we have.
- We also had an estimate column, which has been used to compare our estimation with real estimations.
- In the classification approach we divided the datas in 3 classes and the algorithm works better classifying the first and the last class, this means that highlights well the big difference in price. It has been done using a LDA and trying to use GDA techiniques.
- Finally, to test the real verify the applicability of the model, we splitted the dataframe in train and test set, trained the model on the train and evaluating accuracy on the test set, reaching the results reported above.

Particular mathematical and statistical tools

- In the first part, we performed Shapiro-Francia test to try to prove that datas follow a normal distribution (unsuccessfully).
- The logatithmic rescaling has been widely used in all the project, it's due to the fact that there are very-expensive artworks and if we didn't consider some rescaling, there would be too many outliers.
- In the classification part we used the function SMOTE, from the DMwR library, to oversample the datas in order to better train our model.
- The representations have been performed with some ggplots to show the clusters.

Conclusions

Dealing with this datas and doing researches about topics related to the estimation of art prices, we deeply understand that in a world where everyone try to fix a price and to set evaluation criteria, even if we can set general evaluation standards the personal side and the sensations cannot be mapped.

L'opera d'arte è un messaggio fondamentalmente ambiguo, una pluralità di significati che convivono in un solo significante.
(Umberto Eco)

The work of art is a fundamentally ambiguous message, a plurality of meanings that coexist in a single signifier.
(Umberto Eco)

Thanks for the attention