Project Report: Team 4
Predicting the Price of Art at Auction

Bhavya Agarwal, Farhan Ali, Prema Kolli, and Xiufeng Yang

Department of Computer Science, Stony Brook University,
Stony Brook, NY 11794-4400
{bhavya.agarwal,farhan.ali}@stonybrook.edu,{pkolli,xiuyang}@cs.stonybrook.edu
http://www.cs.stonybrook.edu/~skiena/591/projects

1 Challenge

An art auction is the sale of art works in a competition setting. Our primary goal is to predict the price of an art piece/painting in an art auction. We are focusing only on auctions, which occur in auction houses such as Sotheby's and Christie's. In this project, we try to predict the price of art to be sold at Sotheby's auction in London on 3 December 2014. We are trying to build a generic model, capable of predicting price of any piece of art to be sold based on a variety of features associated with an artist, auction house and the piece of art itself.

There is a famous quote in the movie "The Monuments Men" where the actor George Clooney talks about the importance/significance of art. He says that wiping out any generation's art history is equivalent to wiping out the entire generation itself.

"You can wipe out the entire generation, you can burn their homes to the ground and somehow they will still find their way back. But you can destroy their history, destroy their achievements and this is it they never existed. That is what we are exactly fighting for."

Predicting the price of a painting in an art auction turns out be a challenging task as there are lot of functional dynamics involved in the whole auction process. One challenge in this project is to get such kind of data to build the model and predict the results. Pre-sales estimates are not publicly available to scrape from every website other than Sotheby's and Christie's. But they do not have data that goes back several years, so they expose very limited data. When we look at other sites, pre-sales estimates are only available upon subscribing to their site and at a certain cost. Because of this, it turns out to be a challenging task to get more data with relevant features.

Figures 1,2 and 3 below are showing some of the most Expensive Paintings in the world [5].

Fig. 1. The Scream
Artist: Edvard Munch
Price: $119.9M (123.7M now)
Sold Date: May 2012

Fig. 2. The Card Players
Artist: Paul Cezanne
Price: $259+M (273M now)
Sold Date: April 2011

Fig. 3. Vase with Fifteen Sunflowers
Artist: Vincent van Gogh
Price: $39.7M (83.9M now)
Sold Date: March 1987

2 History/Background

The word 'auction' is derived from the Latin word augeo, which means 'I increase' or 'I augment' [2]. If we look at the history, auctions have been a relatively common way to negotiate the exchange of good and commodities. The first known art auction house in the world was the Stockholm Auction House, Sweden in 1674. Sotheby's currently the world second largest auction house was founded in London in 1744 when Samuel Baker auctioned "several hundred scarce and valuable books" from the library of the Rt Hon Sir John Stanley for a few hundred pounds [3]. Christie's, now the world's largest auction house, was founded by James Christie in London in 1766 and published its first auction catalog in the same year [4].

Auction Process. Some quick insight into the general auction process is given below:

1. Seller request estimate: Estimates are given for all lots and can be based on prices recently paid at auction for comparable property. They can take into account rarity, condition, quality and provenance [1].
2. Auction house listing on catalog: All these auction houses have their catalogs published which are a great way of discovering what's coming up in their future auctions. An auction catalog, that lists the art works to be sold, is written and made available well before the auction date.
3. Exhibition: Also there will be exhibitions which are mostly free and they are a great way for both buyers, sellers and auction houses to exhibit their art and increase their sales.

4. Bidder registration: Persons willing to participate in the bidding/auction process are required to fill out a bidder registration form. If you have not previously bid or consigned with the corresponding auction houses, bidders are required to bring the following [1].
   (a) Individuals: Government-issued photo identification document and proof of current address
   Corporate Clients: a certificate of incorporation
   (b) A recent financial bank statement showing sufficient funds so that you are qualified to bid in the first place.
   (c) If you are bidding on someone's behalf, you have to bring their identification documents too along with yours.
5. Bid (on site, online, by telephone, or written): There are several ways you can bid. It's not necessary to be present on site where the auction is happening. You can bid online or by making a telephone call or even write to the auction houses and place your bid.
6. Pay (price, commissions, buyer's premium): Once you win the bid, you will pay the hammer's price plus buyer's premium on each lot. Buyers premium on each lot varies for every auction house.
7. Delivery: Once you have paid for your lot you can take it home or arrange for it to be delivered.

3 Literature Review

Research on art auction in recent years is mostly concerned with optimal auction strategies for both bidders and sellers in an auction. However in the past, research had been conducted on price formation at traditional auctions and art's potential as an investment. We studied some these research works conducted on modeling art auction, the reliability of estimates and the potential of art as an investment for our literature review.

3.1 Interaction with Domain Expert

To better understand about the intricacies of art valuation, we reached out to Judith Levy, who is the director of Gallery North. Gallery north is a non-profit gallery house here in Long Island, which presents original exhibitions of contemporary art. Judith, who recently took an art valuation course at Christies, was very eager to share her knowledge about the domain and answer our questions.

From our interaction with Judith, we learn that valuing a piece of art was not a simple matter but depended on several factors. She stressed on the importance of factors such as artist popularity, medium of a painting, the size, the category of artwork (contemporary art or 19th century art) and the provenance.

Apart from this, she also mentioned about few of the tricks employed by the auction houses to boost the price of art. For example, some auction houses strategically place valuable paintings in certain lots so as to attract the best bids. Sometimes the auction houses store away valuable paintings if there is not enough demand for the paintings and sell them some time later in the future so as to get optimum price for them.

3.2 Relevant Technical Papers

3.2.1 Modeling On-Line Art Auction Dynamics Using Functional Data Analysis

The first paper we looked at 'Modeling On-Line Art Auction Dynamics Using Functional Data Analysis' [6] by Srinivas Reddy and Mayukh Dass examine the price dynamics at an online art auction of modern Indian art using functional data analysis. Analysis of hedonic products such as art is both complex and challenging. The value of these products depends more on the subjective private value to the bidders than an objective common value. Also the uniqueness of each art work, the scarcity of the art objects along with genre differences among the artists provide a modeling challenge since there is high variability in the data. This paper tries to analyze and identify various factors that can affect the price movement of an artwork at an auction.

Some of the factors identified by the researchers include the artist characteristics (emerging or established artist, prior sales history), art characteristics (painting size, medium used) and competition characteristics such as the number of bidders and pre-auction estimates. They obtained these data for each painting sold at an online auction for modern Indian art in a 3-day auction in December 2004. Then using functional data analysis, they analyze the price velocity and acceleration in on-line auctions. From their results they found that:

1. Established artists show a positive relationship with price at the beginning of an auction
2. For established artist, the rate of price change increases at the end of the auctions
3. Past value of an artists work affect the price level positively in the beginning of the auction
4. Medium of painting (canvas or paper) do not show any relationship with price
5. Size of the painting is negatively correlated with current price during the early part of an auction
6. The number of bidders does not show a significant affect on the price level. But the directionality of the relationship suggests positive influence of the number of bidders on the price levels at the beginning of the auction, showing a decline towards the end.

We found this paper to be very interesting and useful to understand the variables, which can affect the price of an artwork at an auction. For our problem, a functional data analysis method is not suitable since we are trying to predict the final price of an art piece at an auction, but we believe we can use these variables to model the final price of a painting at an auction.

3.2.2 Do Art Specialists Form Unbiased Pre Sale Estimates? An application for Picasso paintings

The second paper we studied was 'Do Art Specialists Form Unbiased Pre Sale Estimates? An application for Picasso paintings' by Corina Czujack and

Maria Fraga O. Martins [7]. This study investigates whether pre sale estimates, provided by the art specialists, are good predictors for the actual price of Picasso paintings sold at Christies and Sotheby's. Christie's and Sotheby's have dominated the art market since the 18th century. They provide high and low pre sale estimates for each artwork they sell. In house art experts provide the pre sales estimates and it depends on several factors such as artist's previous popularity and demand for the genre of artwork.

Pre-sales estimates should be unbiased because to attract a seller, an auction house cannot provide too low an estimate but at the same time, high pre-sale estimates will not attract the buyers. So to study, if a pre-sale estimate is biased at Christie's and Sotheby's, the researchers propose a simple econometric model that models the relationship between the actual selling price and the mid value of the pre sales estimate.

To test whether the pre-sale estimates were unbiased, three different hypothesis were tested. First, the null hypothesis that both Christie's and Sotheby's behave identically in predicting the pricing. Second, if each house releases unbiased prediction and finally the joint hypothesis that both of them behave identically. A Wald's statistics test showed that all the hypothesis could be accepted. This study proves that pre-sales estimates are unbiased and actually good predictors of the actual price of the painting.

3.2.3 Hedonic Models and Pre Auction Estimates: Abstract art revisited

Robert Sproule et al also study the pre-sales estimates. In their paper 'Hedonic Models and Pre Auction Estimates: Abstract art revisited' [8], they compare the predictive power of pre sales estimates to a regression model. The regression model contains various variables such as reputation of an artist, medium, auction house etc. From their research, Sproule et al found that regression models add little or no predictive power above and beyond that of the pre-auction estimates

So previous research has shown than pre-sales estimates are not only unbiased but have strong predictors of the final price of a painting. For our baseline model, we think a simple regression on pre-sales estimate can give us an accurate final price. Also we can utilize pre-sales estimate as a variable in our machine learning models.

3.2.4 Unnatural value: Or art investment as floating crap game

The fourth paper 'Unnatural value: Or art investment as floating crap game' by William Baunmol[9] analyzes the return on investment of artworks. Baunmol studies sales record of paintings over a 300-year period by deceased artists. This is interesting because the supply of artwork is limited and owners of these artworks have complete monopoly on the market. So one would naturally assume the rate of return on these art objects would be better compared to investment in classical securities such as stock or bonds.

But Baunmol concludes that on average the real rate of return on art objects very close to zero. One reason provided by Baunmol for this result was the

fickleness of a buyer's taste. Paintings that command a premium today were at some point considered to be ordinary. So as mentioned previously, this confirms that art is a hedonistic product and its value is derived more from aesthetic pleasure than any objective value.

3.3 BBC Docmentary

We were highly recommended to watch the BBC Documentary on "What makes art valuable?"[14] by the director of Gallery North. Art critic and journalist Alastair Sooke explores the stories behind the Top Ten Most Valuable Paintings in the World to sell at auction. Basically this documentary tells us the stories behind the high prices of art and is a guided tour of the collectors, locales, and Christie's and Sotheby's auction houses that link these great pieces of art together and share some interesting insights.

One of the interesting aspect covered in this documentary is about the provenance of paintings. Provenance of painting basically gives us the chronology of the ownership of that painting. The quality of provenance of painting can make a considerable difference to its selling price in the market and this again is affected by several other factors like the status of past owners, length of stay with a particular owner etc.

To give a simple example, a painting previously owned by David Rockefeller can and does fetch considerably more selling price than another comparable painting without the same ownership record. Picking up a statement from Arne Glimcher who is a renowned art dealer [15], he states that:

"The whole thing of art and money is ridiculous. The value of a painting at auction is not necessarily the value of a painting. It is the value of two people bidding against each other, because they really want the painting."

This was an interesting learning experience but we cannot incorporate provenance information in our prediction model as this information is not available for many paintings and also is very hard to scrape the provenance information from the websites.

4 Data Sets

Obtaining the data for our model was our biggest challenge. We did a thorough search online on art auction databases but most of these databases were not public and were only available for use at a high price. So for our initial baseline model, we decided to obtain data from Sothebys website by scraping the data. Sothebys website provides information about artists name, title of the painting, final price and auction house estimates. Figure 4 below shows us a painting on Sothebys website:

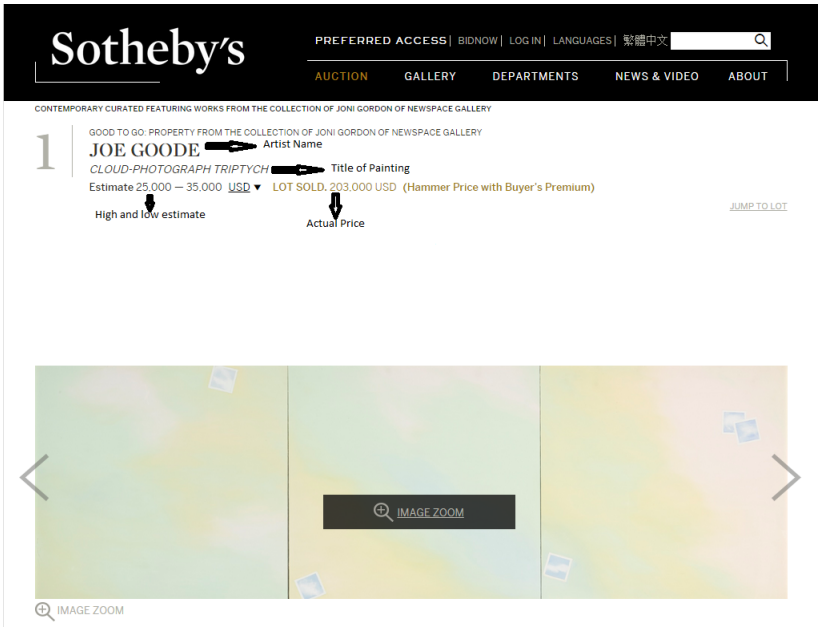


Fig. 4. Sotheby's Website

We obtained the following data for each painting:

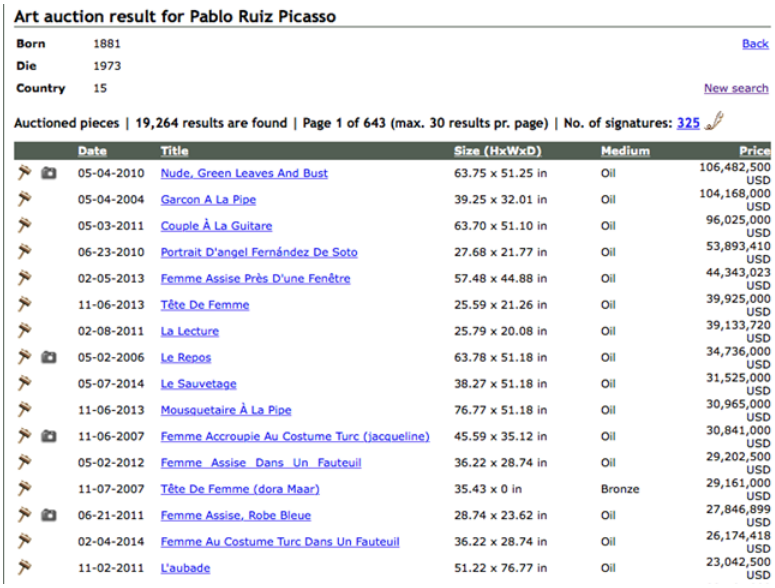
1. Name of the artist
2. Title of the painting
3. Upper estimate
4. Lower Estimate
5. Currency
6. Price it was sold

We also wanted to get paintings from various genres, so we scraped painting from three genres - Contemporary Art, Old Master Paintings and Impressionists and Modern Art. But this proved to be a small data set of 800 records, and we

needed more data to build a satisfactory model. Obtaining more data from the Sothebys website was a challenge because the site did not expose any data going back few years.

The website www.findartinfo.com is an online database of paintings sold at auctions, but just like all the other online databases it did not expose all its information to the public. However it did publish some basic information about a painting such as the title of the painting, artist, medium, date of the auction and the price it was sold for.

Figure 5 below shows us a web page from the findartinfo.com website.



Art auction result for Pablo Ruiz Picasso

Born 1881 Back

Die 1973

Country 15 New search

Auctioned pieces | 19,264 results are found | Page 1 of 643 (max. 30 results pr. page) | No. of signatures: 325

Date	Title	Size (HxWxD)	Medium	Price
05-04-2010	Nude, Green Leaves And Bust	63.75 x 51.25 in	Oil	106,482,500 USD
05-04-2004	Garcon A La Pipe	39.25 x 32.01 in	Oil	104,168,000 USD
05-03-2011	Couple À La Guitare	63.70 x 51.10 in	Oil	96,025,000 USD
06-23-2010	Portrait D'angel Fernández De Soto	27.68 x 21.77 in	Oil	53,893,410 USD
02-05-2013	Femme Assise Près D'une Fenêtre	57.48 x 44.88 in	Oil	44,343,023 USD
11-06-2013	Tête De Femme	25.59 x 21.26 in	Oil	39,925,000 USD
02-08-2011	La Lecture	25.79 x 20.08 in	Oil	39,133,720 USD
05-02-2006	Le Repos	63.78 x 51.18 in	Oil	34,736,000 USD
05-07-2014	Le Sauvetage	38.27 x 51.18 in	Oil	31,525,000 USD
11-06-2013	Mousquetaire À La Pipe	76.77 x 51.18 in	Oil	30,965,000 USD
11-06-2007	Femme Accroupie Au Costume Turc (jacqueline)	45.59 x 35.12 in	Oil	30,841,000 USD
05-02-2012	Femme Assise Dans Un Fauteuil	36.22 x 28.74 in	Oil	29,202,500 USD
11-07-2007	Tête De Femme (dora Maar)	35.43 x 0 in	Bronze	29,161,000 USD
06-21-2011	Femme Assise, Robe Bleue	28.74 x 23.62 in	Oil	27,846,899 USD
02-04-2014	Femme Au Costume Turc Dans Un Fauteuil	36.22 x 28.74 in	Oil	26,174,418 USD
11-02-2011	L'aubade	51.22 x 76.77 in	Oil	23,042,500 USD

Fig. 5. Findartinfo.com's Webpage

From findartinfo.com, we obtained more than 200,000 records of paintings. So our initial data matrix contained 200,000 rows and 8 columns, which are the title of a painting, artist name, birth year of the artist, death year of the artist, date of the auction, dimension of the painting, medium and the price it was sold.

Though we had a huge data set, the data set contained certain number of paintings, which were unsold or below a minimum threshold value. Since our objective was to predict the price of a painting at Sothebys we wanted to filter out cheap paintings, i.e, paintings which were below 10,000 USD (our threshold). We also wanted to filter out all the paintings, which were unsold since it would reduce the noise in the data.

From our initial set of columns, we created few more columns to provide more features for a painting. Features such as aspect ratio, and area were calculated from the dimensions of a painting. Since we also knew the date of an auction

and the birth and death year of an artist, we wanted to use this information to calculate the difference between the auction year and birth year and auction year and death year. Another feature we added was the average price of paintings sold by an artist, though we did not consider the year at which the painting was sold when calculating the average.

Table 1 below shows the summary of the data matrix:

Initial Number of rows	200,000
After removing unsold paintings	125,000
After removing cheap paintings (below 10K)	25,000
Number of columns	10

Table 1. Dataset Statistics for Baseline Model

The columns in the data matrix are:

1. Medium of the painting
2. Width of the painting
3. Height of the painting
4. Area
5. Aspect Ratio
6. Artists average price
7. Year the painting was sold
8. Number of years since the artist was born
9. Number of years since the artist died
10. Price of the painting (inflation adjusted)

We were still not satisifed with 25,000 records we had in our dataset. We decided to look at paintings which are above $1,000,000 and discard the rest. So we went ahead and found all those artists whose works have been sold for more than $1,000,000 and we got a list of 700 artists. We went back to findartinfo.com and scraped all the paintings for these 700 artists in our list. Now we have our final rich dataset. We wanted to get the images for these paintings as well as it would be important to include image-related features in our final advanced model. However findartinfo.com had images for only about 18,000 paintings. After a lot of technical difficulties, we were able to scrape the images for the paintings and we got around 18,000 images. Finally we have all the data we need, all the paintings which were sold from year 2002 to year 2014 for our 700 artists.

Table 2 below shows the summary of **our final data matrix:**

Initial Number of Paintings	50,000
After Cleaning	40,000
Number of Images	18,000
Number of Data Fields	23
Number of artists	700

Table 2. Dataset Statistics for Intermediate Model

Let us begin by looking at the features for Paintings:

1. **Price of the painting (inflation adjusted) :** This is the actual selling price. We adjusted the price to the inflation with the Consumer Price Index inflation data. We brought it all to the common time line of September 2014.
2. **Year the painting was sold :** This feature is self-explanatory.
3. **Width of the painting :** This feature is self-explanatory.
4. **Height of the painting :** This feature is self-explanatory.
5. **Area :** We calculated the area based on the painting dimensions and added it as a feature.
6. **Aspect Ratio :** We calculated the aspect ratio based on the painting dimensions and added it as a feature.
7. **Medium of the painting :** This is a categorical variable. We have categorized the medium into three categories : Oil on Canvas, Paper and Others.
8. **Auction House :** This again is a categorical variable. We have categorized the Auction Houses into three categories : Sotheby's, Christie's and Others.
9. **Auction House Low Estimate :** This feature is self-explanatory.
10. **Auction House High Estimate :** This feature is self-explanatory.
11. **Premium :** This is a binary variable indicating if premium is included in the actual price or not so that we can take care of the calculations accordingly.
12. **Date when the Painting was made :** This feature is self-explanatory.

Now let us look at the features for artists:

1. **Significance :** This feature measures the contribution to the world when compared with the average human being.
2. **Gravitas :** This feature captures the degree of name recognition based on the subject's accomplishments.
3. **Celebrity :** This feature takes into account combination of the page length, the number of edits, and the number of hits for the subject's page as celebrity.
4. **Fame :** This feature combines the Gravitas and Celebrity quotients. Fame is subject to fluctuations over time which is taken care of.
5. **Number of years since the artist was born :** This feature is self-explanatory
6. **Number of years since the artist died :** This feature is self-explanatory

Finally let us look at the features for Images we have in the dataset. For using the visual features of the image, we created four features covering various aspects of the painting, based purely on image processing techniques. These will be described more in section 7.2.

1. **Color Palette Score :** This feature gives us an idea of the richness in terms of number of colors used in the painting.
2. **Is it a Portrait :** This feature was used to discriminate the portraits from rest of the images.
3. **Smoothness Score :** This feature is used to give an indication if the image has lot of objects, or a lot of colourful strokes creating sharp contrast at many points.
4. **Brightness Score :** This feature tells us if the image has a lot of bright colors, or it has predominantly darker shades to it.

5 Observations

We have learnt a couple of interesting things from our data set. Please find those observations along with the plots in the subsections below.

5.1 Year Wise Trend

We wanted to look at the year wise trend of how the painting's selling prices vary with respect to auction house estimates in terms of whether they are in range, lower or higher than the auction house estimates. Looking at the plot below in Figure 6, we can see that there is a clear trend with economic conditions between the years 2002 and 2014. It can be clearly seen that the paintings were exceeding the estimates at the time of the economic boom in the early 21st century. However, with the crisis of 2008, the ratio fell down drastically and is still recovering year by year, just like the economic market.

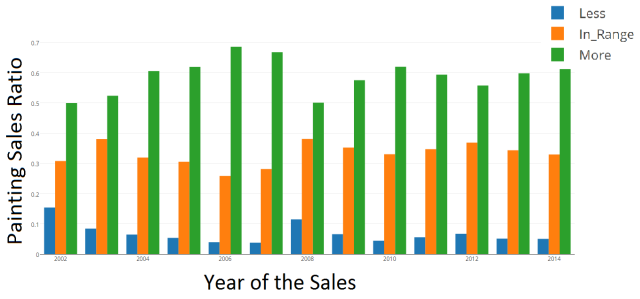


Fig. 6. Year Wise Trend

5.2 How accurate are Auction House Estimates?

We wanted to understand how accurate the auction house estimates are. As we were targeting Sotheby's and Christie's, we wanted to understand which of these two auction houses did a better job in coming up with the estimates closer to the actual selling price. Sotheby's did a much better job at the estimates when compared to Christie's. Figures 7 and 8 below show the error plots for both the auction houses.

Auction House/Error (%)	Median Error (%)	Mean Error (%)
Sotheby's	17.63	25.74
Christie's	23.64	26.26

Table 3. Auction House Estimates Error Percentages

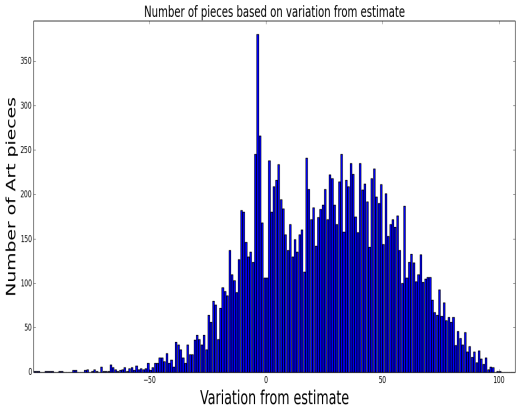


Fig. 7. Sotheby's Error Percentages Plot

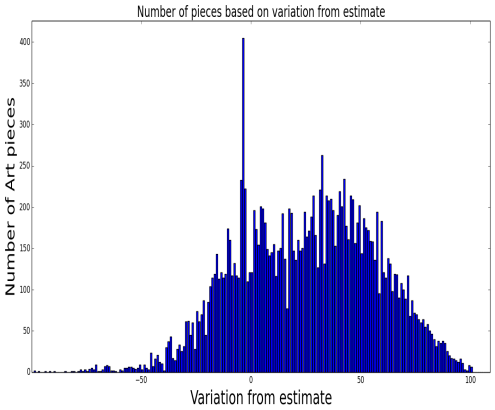


Fig. 8. Christie's Error Percentages Plot

5.3 Artist's Career Peak

We wanted to understand the career highs and lows of an artist. So we picked top three artists in our dataset with highest number of paintings. They were Marc Chagall, Pablo Picasso and Andy Warhol. When we look at their career graphs, as shown in Figures 9, 10 and 11, in terms of how many painting's they have painted against the price they have been sold for, we see that there was a peak phase in Andy Warhol's career during the year 1964 unlike Marc Chagall and Pablo Picasso. Their career's were pretty consistent throughout except for a spike in one or two paintings here and there during their life span.

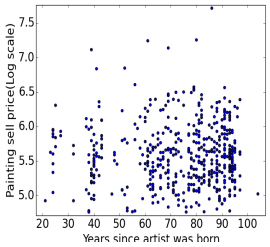


Fig. 9. Marc Chagall's Career Graph

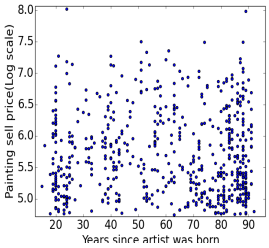


Fig. 10. Pablo Picasso's Career Graph

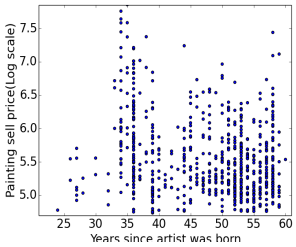


Fig. 11. Andy Warhol's Career Graph

6 Baseline Model

In this section, we have briefly explained our baseline model along with initial prediction results and we have verified the initial prediction results as well after that specific lot was sold at Sotheby's auction. Please find more details below.

Before jumping on to the model, we first plotted a graph depicting the number of paintings with actual selling price lower or higher than the estimates given by auction house. The auction house estimates proved pretty accurate for most of the paintings as shown in Figure 12. These estimates are based on prices recently paid at auction for comparable property. They can take into account rarity, condition, quality and provenance. We then started off with a simple linear

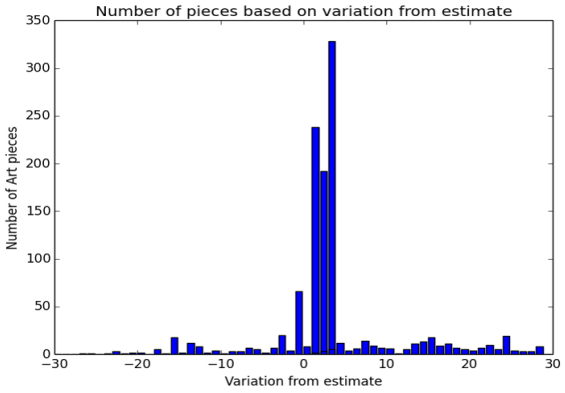


Fig. 12. Variation of Actual Price from Auction House Estimates

regression model, taking the mean of the lower and upper estimates given by the auction house as a feature for prediction. We used ElasticNet, which reduces overfitting and obtained a line which fit quite well with the actual data. We obtained a r2 score of about 0.75 (with best being 1). We think that this is pretty reasonable for a baseline model. We got the coefficient of 1.45 and an intercept of -3030. This shows that the model predicts the actual price to be higher than the auction houses estimates, as the intercept is seemingly insignificant here. This was in line with the observations in the previous paragraph. The plot we obtained is shown in Figure 13. After this we also tried another slight variation of baseline model taking the upper and lower auction house estimates as different features. However it did not yield any better results, as both the estimates tend to be in a certain proportion of each other. We plotted how much our estimations varied from the actual value, in terms of percentages and found that the model performs worse than the previous one, with an r2 score of -0.5. The coefficients were 10.89 for higher estimate and -6.23 for the lower estimate along with an intercept of -11,756. The results were worse than the previous model, indicating

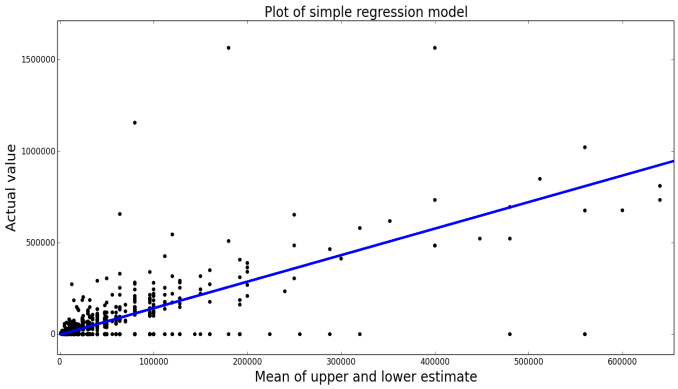


Fig. 13. Line fitted by Linear Regression for the Baseline Model

that using both the values possibly confuses the predictor. As we can see, most predictions varied more than 50% from the actual value which is not an accurate measure. The plot of variation is as shown in figure 14. Since we were having

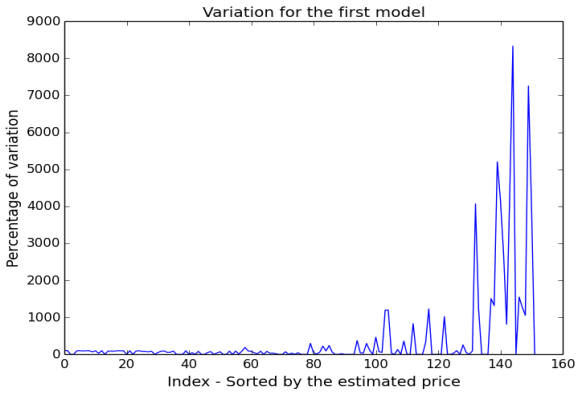


Fig. 14. Variation of Baseline Model's Performance

very low number of features/data fields, we did not tend to go beyond linear regression. Also initial baseline model gave us decent results to start with by using the mean of the upper and lower auction house estimates as a feature for the linear regression model.

We used this baseline model to generate a prediction for a painting at October 17 auction at Sotheby's London. Figure 15 is showing us Adrian Ghenie painting titled Duchamp's Funeral I. This painting was sold for $1.33 Million, while we

predicted the price to be $1.2 Million which is closer than the auction house prediction of $640,000-$960,000.

Fig. 15. Adrian Ghenie's - Duchamp's Funeral I

7 Advanced Model

7.1 Data Preprocessing/Cleaning

First step was to filter the data. We removed the paintings which were made on lithograph, and then removed all the 3-D sculptures from the data. Since we were concentrating on only the paintings sold over 100K, we had most of the fields available and the data didn't require much cleaning. After that, we first adjusted the currencies based on the exchange rates for the year the painting was sold. We got the data from "grandtrunk" api which gave us the currency rates for last 15 years. After that we adjusted the price to the inflation with the Consumer Price Index inflation data. We brought it all to the common time-line of September 2014 (Note: we removed the cheaper paintings after adjusting the price).

7.2 Feature Construction

1. **Image Features**
   (a) Color Palette Score : For this we used k-means color quantization to reduce the colors in an image. In this algorithm, we use the K-means algorithm to cluster pixels based on similar colors and reduce the total number of colors to a pre-defined number, i.e. the number of clusters k as passed in the k-means initialization. We set k = 5, i.e. we reduced the number of colors in the image to 5. After this, we subtracted the

source image from the new image, and took a mean across the intensity changes in all the pixels. Hence, if the image was more colourful, the feature would have a higher value as compared to one which is more monotonic in nature, thus making it possible to club the more colourful images together.

(b) Is Portrait : These type of images were the easiest to separate, as there are many state-of-the-art methods available in open-source libraries like OpenCV. For this case, we used a pre-trained Haar-cascade classifier for face detection in the image. The faces in the recent images tend to be more abstract, and hence are more challenging to detect than the standard images used in computer vision. Hence, we also used the Haar-cascade for eye detection, and returned a confidence value of 0.5, if there was no face detected but there were eyes in the image. This made our feature more robust to abstract paintings.

(c) Smoothness Score : For calculating this, we ran the Canny edge detector on the image, and found out the number of pixels in the image corresponding to an edge. For the final score we simply divided the number of edge pixels to number of total pixels in the image to obtain a normalized score.

(d) Brightness Score : We changed the image color space from RGB to YUV, because Y-channel corresponds to the luminosity of the image. We simply averaged the Y-channel values for each pixel in the image and obtained a brightness score.

2. **Artist's and Painting's Features**

(a) Artist's Significance: For the 700 artist names in our list, we wanted to link it to Prof. Skiena's Fame metric [16], however most of the artists were obscure and also the names from scraping were not in a clean and consistent format. Hence, while our script was able to get the data for most of the artists, we had to manually pick the values for many of them. We used the fields 'gravitas', 'significance', 'fame' and 'celebrity' as they seemed to be most correlated with the artists market price.

(b) We also enumerated the mediums and created a dictionary, where we correspond the medium to a number : Oil on Canvas, Paper and Others.

(c) We changed the birth and death year of the artist to year since he died/born to get a better representation for the predictor.

(d) We also enumerated the auction houses to be of three values : Sotheby's, Christie's and Others.

(e) Finally we extract width and height from dimensions and add 2 new features which are area and aspect ratio of the painting.

Table 4 below is showing us the image features along with their lowest and highest scores.

After constructing image features, we went ahead and looked at if the feature values are accurately representing the image. Fig.16 and Fig.17 are extreme examples of color palette score. While Fig.16 has just a single color, Fig.17 has a

Features/Scores	Low-Score	High-Score	Paintings
Color-Palette Score	0.829	89.3	Figures 16, 17
Is-Portrait	0	1	Figures 18, 19
Smoothness Score	0.000001	0.44	Figures 20, 21
Brightness score	0.0272	0.974	Figures 22, 23

Table 4. Image Features along with their lowest and highest scores

more colors, which heavily contrast with each other giving it a high score. Fig.18 is the example for is_portrait feature failing due to distorted faces due to the abstract nature of the art, while Fig.19 is a textbook example of portrait, and the feature works perfectly for this one. Fig.20 has a low smoothness score due to very low number of high contrast edges, while Fig.21 is very rough to look at and looks like a very ragged surface, giving it a huge number of edges. Finally, Fig.22 is has low brightness score, because the painting is predominantly black, while Fig.23 is completely white, with just 2 small grey squares resulting in high brightness score.

Fig. 16. Painting with Low Color-Palette Score of 0.829

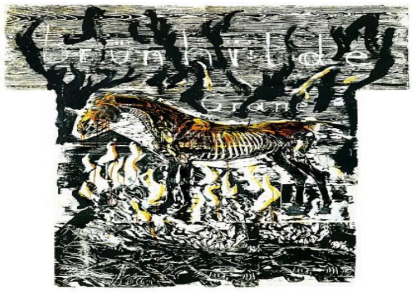


Fig. 17. Painting with High Color-Palette Score of 89.3

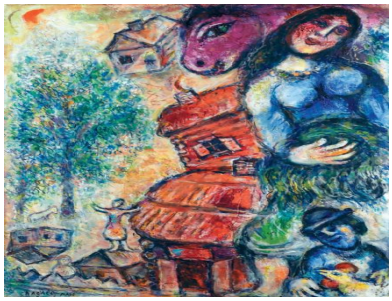


Fig. 18. Painting with Is-Portrait 0

Fig. 19. Painting with Is-Portrait 1

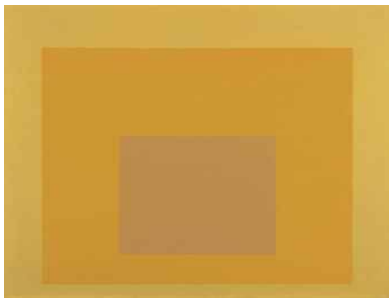


Fig. 20. Painting with Low Smoothness Score of 0.000001

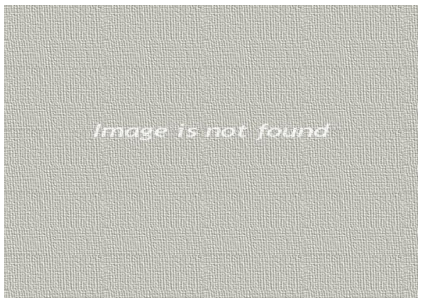


Fig. 21. Painting with High Smoothness Score of 0.44

Fig. 22. Painting with Low Brightness Score of 0.0272

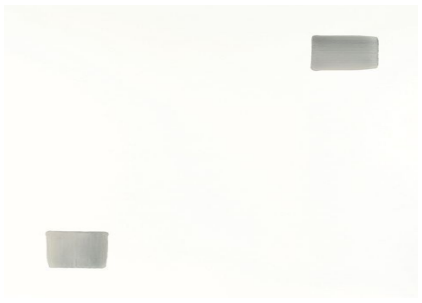


Fig. 23. Painting with High Brightness Score of 0.974

7.3 Correlation

After pre-processing the data and constructing features, first thing we wanted to do before jumping on to building the model is to have a look at how the features are correlated with each other and also with the actual selling price which is our output variable. We did not include the image features in the below correlation

matrix as we do not have them for all the 40K paintings. We have them only for around 18K paintings. Please find the correlation for artist's and painting's features in Figure 24.

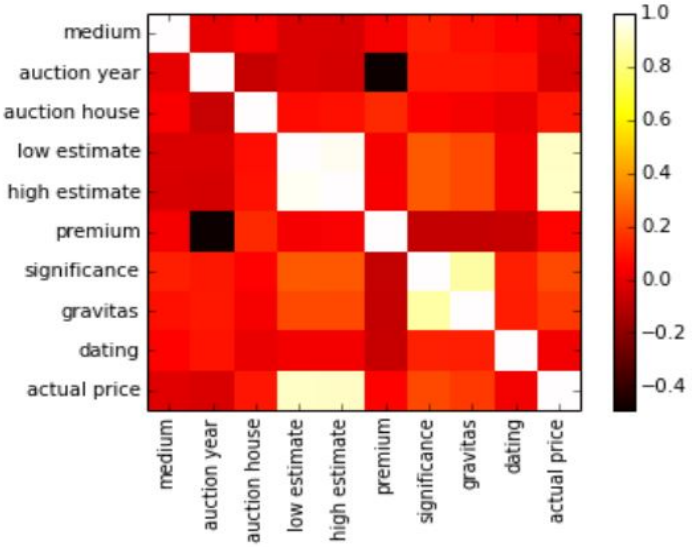


Fig. 24. Correlation Heat Map for Features (excluding image features)

We were curious to see how four of our image features were correlated with the actual price. It turns out that only 'Is Portrait' has a correlation of about 0.06 which affected our model positively. Brightness and Smoothness score both are negatively correlated and way less than 0.05. Color Palette score correlation was 0.03. So still there is huge scope for constructing better image features in the future now that we have good enough number of images for these paintings in the dataset. Please find the correlation map in Figure 25.

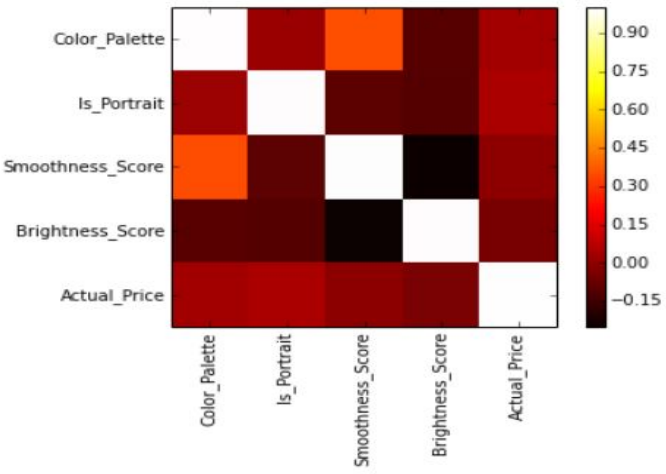


Fig. 25. Correlation Heat Map for Image Features

Finally these are the features that were actually correlated to the selling price of the painting. Please find the list of significant features in the order of significance from highest to lowest below.

1. Lower Auction House Estimate
2. Upper Auction House Estimate
3. Medium
4. Significance
5. Height
6. Width
7. Gravitas
8. Fame
9. Celebrity
10. Auction House
11. Premium

7.4 Building the model

This time we tried 3 different models : Linear Regression, K-Nearest Neighbors Regression and Decision Tree Regression. The actual selling price and the estimates both follow the log-polar distribution. Hence, we first took the log of both the values. We then extrapolate the predicted values with anti-log before making the prediction.

7.4.1 Linear Regression

In linear regression, data is modeled using linear predictor functions. We try to fit linear coefficients to the features, by trying to minimize the least squares

error function. We split the data into 80:20 ratio for training and validation respectively. After training the predictor, we got an average absolute error of around 35% and average absolute median error of 30%. That implies that more than half of the predictions were in the range of 25% of the actual selling price. This is a great improvement over our last models and shows the importance of auction house estimates, and the effect they have. Here is the graph (as shown in Figure 26) that we obtain for the linear regression model.

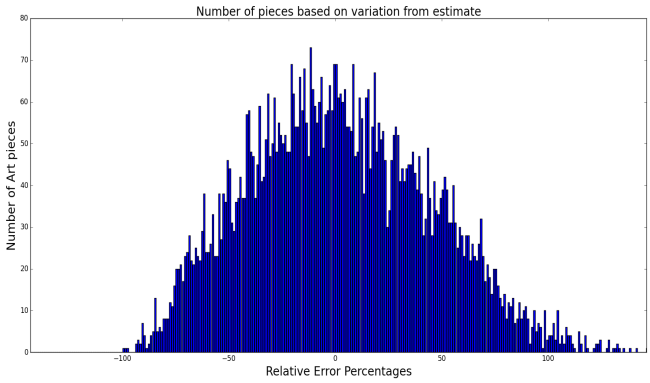


Fig. 26. Performance of Linear Regression Model

7.4.2 k-Nearest Neigbors Regressor

After that we thought that k-NN Regressor would be a better predictor for this problem. This predictor considers only the K-nearest neighbors while training the predictor for an example, which captures the relation between similar paintings. This is more intuitive as we would expect similar paintings to have more effect on the prices. However, we were proved wrong, and as it turned out, the global properties also have a huge effect on the painting's price, and a small number of neighbors is too less a number to accurately estimate a painting's price. We found the optimal 'k' value to be 5 after performing cross-validation and obtained the following plot. We got a median error of 61.3% and mean absolute error at 102.8%. Here is the graph (as shown in Figure 27) that we obtain for the k-NN regression model.

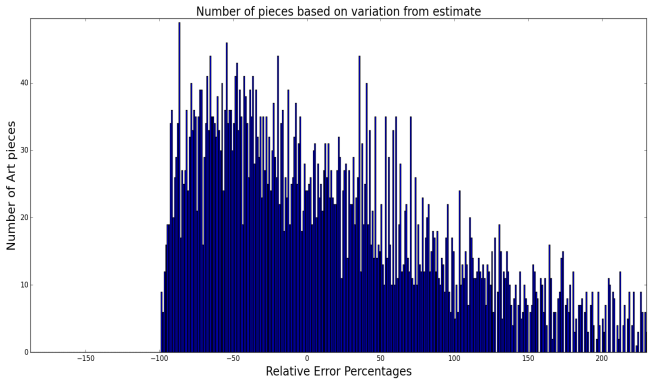


Fig. 27. Performance of k-NN Regression Model

For k-NN regression model, we tried to give just the image features to the model, to see how it performed. And we tried to look at the nearest neighbors for couple of test instances. To be more specific, we looked at what are the nearest images for a test image that we are regressing on, and the neighbors looked absurd and totally not similar. It did not do a good job with the image features. Basically from the correlation analysis we know that these image features are not statistically significant, so it is boiling down to constructing image features which are significant and then probably we can expect k-NN to do a better job at that.

7.4.3 Decision Tree Regressor

The decision tree was a new choice considering the features like auction house, medium etc. which have only few enumerated values. A decision tree classifier is perfect for making sub classifiers depending on the painting's properties. Also, in this case the final classifier can concentrate on the paintings which belong to a

certain category, and hence can exploit the common properties they have. While the decision tree performed better than k-NN, considering it had a lot more neighbors in the sub-classifiers, it still did not match the accuracy of Linear Regression, however it came pretty close. The predictions had a median error of 34.8% and mean absolute error at 52.4%. Here is the graph (as shown in Figure 28) that we obtain for the decision tree regression model.

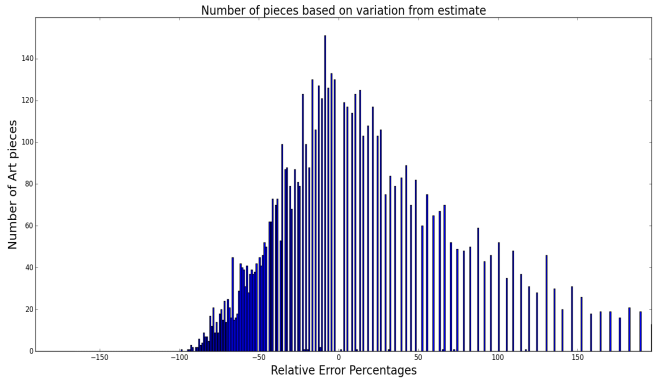


Fig. 28. Performance of Decision Tree Regression Model

Table 5 below is showing us the comparison between all the advanced models without image features for the entire dataset. Linear Regression model performed the best.

Model/Error(%)	Median Error (%)	Mean Error (%)
Linear Regression	30.8	35.4
Decision Tree	34.8	52.4
k-NN Regressor	61.3	102.8

Table 5. Comparison of all advanced models without image features for the entire dataset

7.4.3 Comparison of all three models with and without image features

Now since we obtained images for only about 18,000 paintings, we are including a separate subsection here for comparing the performance of these models with and without image features for just these 18,000 records and not for the entire dataset.

Table 6 clearly shows us that Linear Regression and Decision Tree Regressor did slightly better with image features. However k-NN Regressor performed better without image features.

Model/Error(%)	Median Error (%)	Mean Error (%)
Linear Regression (with Image Features)	32.56	38.9
Decision Tree (with Image Features)	32.39	56.63
k-NN Regressor (with Image Features)	54	63.67
Linear Regression (without Image Features)	32.67	38.96
Decision Tree (without Image Features)	34.89	58.22
k-NN Regressor (without Image Features)	38.13	45.21

Table 6. Comparison of all three models with and without image features for just 18K records

Figures 29 and 30 are showing us the Performance of Linear Regression Model with and without image features respectively. As you can see it did slightly better with image features.

Figures 31 and 32 are showing us the Performance of Decision Tree Regression Model with and without image features respectively. As you can see it did slightly better with image features.

Figures 33 and 34 are showing us the Performance of k-NN Regression Model with and without image features respectively. As you can see it did worse with image features. Distance function here is not doing a good job at all when image features are included.

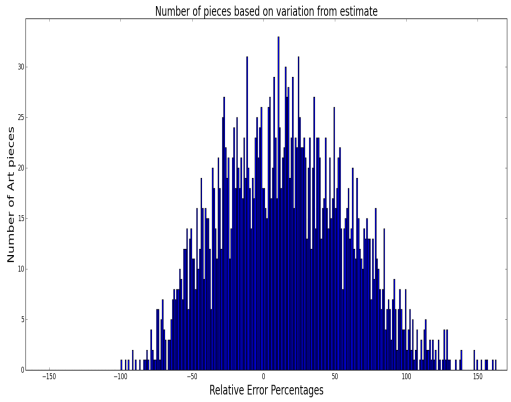


Fig. 29. Performance of Linear Regression with Image Features

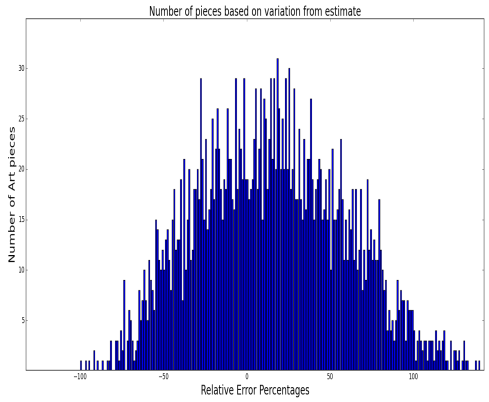


Fig. 30. Performance of Linear Regression without Image Features

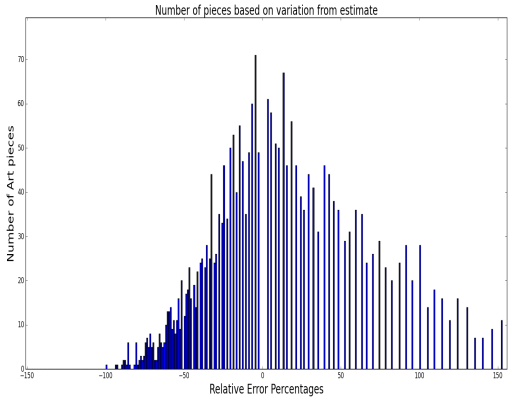


Fig. 31. Performance of Decision Tree with Image Features

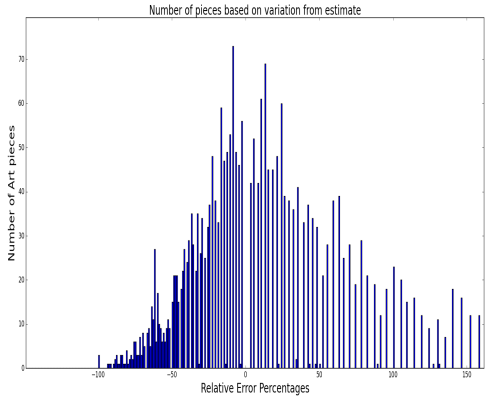


Fig. 32. Performance of Decision Tree without Image Features

8 Final Prediction and Conclusions

For our final prediction, we picked the Old Master and British Paintings auction which happened in Sotheby's, London on 03 Dec 2014. There were 16 lots at the auction and these were the paintings of 12 famous artists. Two of these lots were unsold.

Please find our predictions in Table 7 below:

Artist	Painting Title	Low_Est	Upp_Est	Mid_Est	Dec_Tree	Lin_Reg	Sold_Price
David Teniers (ii)	a countryside inn with revellers enjoying an afternoon's drinking, the innkeeper chopping wood in the foreground, and a pastoral landscape	100,000	150,000	125,000	157101.3706	215971.7316	422500
Jan Brueghel	a winter landscape with travellers passing through a village	100,000	150,000	125,000	157101.3706	211060.7162	386500
Willem Van De Velde	a calm sea with a kaag and a boeier close in to the shore, other ships beyond	300,000	400,000	350,000	991242.6361	538503.9913	722500
Claude Joseph Vernet	a mediterranean coastal scene at sunset with figures fishing in the foreground	300,000	500,000	400,000	320759.8295	524803.069	362500
Giovanni Antonio Canal Canaletto	venice, the piazza san marco looking east towards the basilica	5,000,000	7,000,000	6,000,000	4228438.067	6222005.859	5458500
Artemisia Gentileschi	bathsheba at her bath	200,000	300,000	250,000	583686.9372	422991.4709	602500
Jusepe De Ribera	saint teresa of avila	150,000	200,000	175,000	127696.7882	311377.2947	122500
Peter Paul Rubens	the martyrdom of saint paul	600,000	800,000	700,000	640000	968971.0742	914500
Willem Van De Velde	the jupiter and another dutch ship wrecked on a rocky coast in a gale	400,000	600,000	500,000	403812.7005	701065.1702	722500
Adriaen Coorte	three peaches on a stone ledge, with a red admiral butterfly	2,000,000	3,000,000	2,500,000	824479.7131	3219874.425	3442500
Jean Marc Nattier	portrait of francis greville, baron brooke, later 1st earl of warwick	150,000	250,000	200,000	193276.9101	307600.0324	386500
Joseph Mallord William Turner	rome, from mount aventine	15,000,000	20,000,000	17,500,000	32823128.58	14999206.83	30,322,500
David Teniers (ii)	peasants playing nine-pins outside an inn, the city of antwerp in the distance	100,000	150,000	125,000	292536.4414	211809.894	254500
Jan Brueghel	travellers passing through a village in winter	100,000	150,000	125,000	121949.4859	212272.8907	302500

Table 7. Our Predictions along with the actual sold price using Linear Regression and Decision Tree

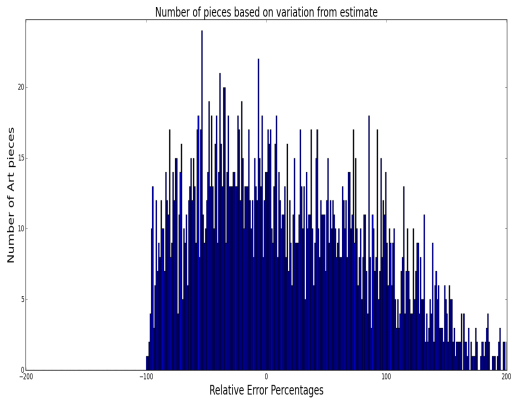


Fig. 33. Performance of k-NN Regressor with Image Features

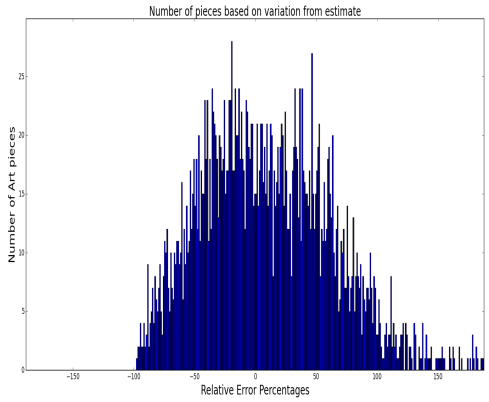


Fig. 34. Performance of k-NN Regressor without Image Features

Looking at the Tables 8 and 9, it is pretty clear that we did very good when compared to the auction house estimates. Linear Regression Model performs better in most of the cases. However Decision Tree Regressor almost nailed it for few paintings. Overall Linear Regression is the best we have got.

Showing two random paintings which were auctioned at the Old Master and British Paintings auction which happened in Sotheby's, London on 03 Dec 2014 below in Figures 31 and 32.

Fig. 35. Title: Rome, From Mount Aventine
Artist: Joseph Mallord William Turner
Our Prediction: 32,823,128.58 GBP
Sold Price: 30,322,500 GBP
Sold Date: 3 Dec 2014

Fig. 36. Title: Venice
Artist: Giovanni Antonio Canal
Our Prediction: 6,222,005.859 GBP
Sold Price: 5,458,500 GBP
Sold Date: 3 Dec 2014

Artist	Error_Low_Est	Error_High_Est	Error_Mid_Est	Error_DecTree	Error_LinReg
David Teniers (ii)	76.33	64.5	70.41	62.8	48.8
Jan Brueghel	74.12	61.2	67.65	59.3	45.4
Willem Van De Velde	58.47	44.6	51.55	37.2	25.4
Claude Joseph Vernet	17.24	37.9	10.34	11.5	44.77
Giovanni Antonio Canal Canaletto	8.4	28.2	9.9	22.53	13.9
Artemisia Gentileschi	66.8	50.2	58.5	3.12	29.8
Jusepe De Ribera	22.4	63.2	42.85	4.24	154.18
Peter Paul Rubens	34.4	12.5	23.45	30.01	5.95
Willem Van De Velde	44.6	16.9	30.8	44.1	2.96
Adriaen Coorte	41.9	12.8	27.37	76.04	6.46
Jean Marc Nattier	61.2	35.31	48.25	49.99	20.41
Joseph Mallord William Turner	50.53	34.04	42.3	8.24	50.53
David Teniers (ii)	60.7	41.06	50.8	14.9	16.77
Jan Brueghel	66.94	50.41	58.6	59.6	29.82

Table 8. Relative Error Percentages with respect to estimates and our models

Model/Error(%)	Median Error (%)	Mean Error (%)
Error_Low_Est	54.5	48.8
Error_High_Est	39.49	39.5
Error_Mid_Est	45.5	42.35
Error_Dec_Tree	33.6	34.5
Error_Lin_Reg	27.6	35.38

Table 9. Mean and Median Error Percentages with respect to estimates and our models

Now to conclude we are confident that we have built a good model as our advanced model has performed better than the estimates provided by the auction houses. However for future work, one can definitely look at constructing more image features which are statistically significant and further, more work can be done on this aspect where one can employ deep image processing techniques to come up with strong features.

References

1. Christie's Buying Guide http://www.christies.com/features/guides/buying/forms/How%20to%20Buy%20NY.pdf
2. Auction http://en.wikipedia.org/wiki/Auction
3. Sotheby's http://www.sothebys.com/en/inside/about-us.html
4. Christie's http://www.christies.com/about-us/company/overview/
5. List of Expensive Paintings http://en.wikipedia.org/wiki/List_of_most_expensive_paintings
6. Reddy, Srinivas K., and Mayukh Dass. "Modeling on-line art auction dynamics using functional data analysis." Statistical Science (2006): 179-193. APA
7. Czujack, Corinna, and Maria Fraga O. Martins. "Do art specialists form unbiased pre-sale estimates? An application for Picasso paintings." Applied Economics Letters 11.4 (2004): 245-249.
8. Baumol, William J. "Unnatural value: or art investment as floating crap game." The American Economic Review (1986): 10-14.
9. Calin Valsan & Robert Sproule, 2006. "Hedonic Models and Pre-Auction Estimates: Abstract Art Revisited," Economics Bulletin, AccessEcon, vol. 26(5), pages 1-10.
10. Afghan Girl http://en.wikipedia.org/wiki/Afghan_Girl
11. Lunch atop a skyscraper http://en.wikipedia.org/wiki/Lunch_atop_a_Skyscraper
12. The Starry Night http://en.wikipedia.org/wiki/The_Starry_Night
13. Mona Lisa http://en.wikipedia.org/wiki/Mona_Lisa
14. BBC Documentary on "What makes art valuable?", https://www.youtube.com/watch?v=QXOPBZFvBQ4&feature=youtu.be
15. Article on "What Makes Art Valuable?", http://topdocumentaryfilms.com/what-makes-art-valuable/
16. Who's Bigger?, http://en.wikipedia.org/wiki/Who's_Bigger%3F