Price Prediction of an Art at Auction

Kashish Kothari
(202118011)
MSc DS, DAIICT
Gandhinagar, Gujarat
kashishk9.9.00@gmail.com

Jainam Shah
(202118014)
MSc DS, DAIICT
Gandhinagar, Gujarat
jainamshah535@gmail.com

Dhvani Golani
(202118020)
MSc DS, DAIICT
Gandhinagar, Gujarat
dhvanigolani2000@gmail.com

Prachi
(202118021)
MSc DS, DAIICT
Gandhinagar, Gujarat
prachik1804shah@gmail.com

***Abstract*—An art auction is the sale of artworks in a business context. Our primary purpose is to forecast the price of an art piece/painting in an art auction. We are interested in auctions that take place online on platforms such as Artsy.net, which unites top auction houses, NGOs, and sellers from across the world in one place. In this research, we attempt to forecast the price of paintings sold at these auctions. We are attempting to develop a generic model capable of forecasting the price of any piece of art to be sold based on various factors connected with an artist, auction house, and the piece itself.**

I. Introduction

Startups focusing on the art world have raised around 228 million dollars in venue capital finance since 2003. Almost 75% of the funding has gone to art start-ups focused on the high-end art market. Startups aimed at the wealthy, also known as the 1%, received more than 73% of the financing allocated to the art market space. One notable example is the high-end, customer-focused luxury art and antique provider 1stdibs, sponsored by Benchmark Capital. It is important to remember that Benchmark does not only support the affluent, but it also supports Art.com, a firm that sells affordable art and is almost ready to go public.

Before the auction, predictions of the painting price are central to the art auction process. Very highly-skilled professionals make predictions with comprehensive training. Due to the time sparse of these experts, any automation would assist us. The cost of hiring such experts is also not economical for small-scale auction houses.

The ability to accurately predict an art item's price is crucial for auction houses and artists. It helps the houses choose which works to group together and display at a specific exhibition, and it supports artists' planning and decision-making between the creation of a piece and its sale. Removing human biases during appraisal, such as projecting high values for artists affiliating with well-known and vital institutions, will help artists who use automation in art price prediction. Automated evaluation that is independent of the artist would be a step in the direction of a more transparent and fair art market. Last but not least, from the viewpoint of the market, more publicly disseminated price forecasts would boost liquidity: more excellent knowledge would encourage more educated risk-taking, which would boost total productivity for both buyers and sellers (Bailey [2020]). To everyone's advantage, automation of prediction would get beyond the capacity bottleneck of valuation professionals, greatly enhancing public access to this vital information.

We intend to propose a project that will accurately predict the prices of paintings using Machine Learning. This project considers all the factors that affect the prediction of an art price and creates a data set that is liable and ideal. This project intends to fabricate a model that predicts the art price using the art piece and significant features before the auction.

II. Data Extraction

Data mining or collection is a very primitive step in the data science life cycle. People may use a massive amount of data on the internet to satisfy their business needs. Some tool or approach is required to acquire this information from the web. Additionally, this is where the concept of web scraping is used.. Web scraping is the method for extracting and creating a structured representation of data from a website.
The website ARTSY, like many others, has its structure, form, and a tonne of readily available user data. However, because it lacks a defined API, obtaining data from the site is challenging. To create our data set, we will web scrape the website to collect the unstructured website data and organize it.

Selenium is a Python library and tool used for automating web browsers to do several tasks. A web driver is a vital ingredient to this process. It is what will automatically open up a browser to access the website of your choice. We open the Artsy website in Google Chrome using Selenium Web Driver.
Web Driver: It automatically opens up the browser to access the website of your choice.
Selenium provides a wide range of ways to interact with sites, such as:

- Clicking Buttons- Click the Login button to log in to the website. Also, for clicking the "More Info" button for every painting detail.
- Populating forms with data- Helps populate email id and password for logging in.
- Scrolling the page- Allows to scroll down to go to the next page.
- Executing your custom code- Finds the HTML element from which data is to be extracted.

Libraries used for Web Scraping:

- Selenium: Selenium is a web testing library. It is used to automate browser activities.
- Pandas: Pandas is a library used for data manipulation and analysis. It is used to extract the data and store it in the desired format.

III. Data Description

Our data set consists of 15 columns and 694 rows.

Column name	Description	Sample
Id	Unique key of each entry	1
Painting Name	Name of painting made by artist	Thames Scene with Power Station
Year	Year When the painting was made	1990
Painting Type	Medium and technique with which the painting was made	oil on canvas
Auction Date	Selling date of painting	12-05-2022
Percent Change	Percent change in sales price and estimated price	10
Painting size length	Length of painting	252.7 (cm)
Painting Size Breadth	Breadth of painting	144.8 (cm)
Estimated price from inDollars	Upper bound of est.price	7000
Estimated price to inDollars	Lower bound of est.price	10000
Sale price inDollars	Selling price of painting	7650
Artist	Artist Name	Tom Wesselmann
Artist's Nationality	Artist's Nationality	American
Auction Name	Auction house name	Sotheby's
Painting links	Link to painting	https://d7hftxdivxxvm.cloudfront.net/?resize_to=fill&width=100&height=100&quality=80&src=https%3A%2F%2Fd32dm0rphc51dk.cloudfront.net%2FZpJJZTr6NgN1JJHqrV6l0Q%2Flarger.jpg

The data consists of nine profound artists, namely:

Artist's Name
Gerhard Richter
Tom Wesselman
Takashi Murakami
Georg Baselitz
Francis Newton Souza
Maqbool Fida Husain
George Condo
Salman Toor
Sayed Haider Raza

IV. Data Pre-Processing

The original data set needs pre-processing. The following are the steps taken to manipulate the data to get better and more efficient results.

1) Splitting of columns: There were several columns that needed to be split into further columns. Painting_size_length and Painting_size_breadth, Estimated_price_from_inDollars and Estimatied_price_to_inDollars and Year were delimited using the excel delimiter function.

2) Converting Price to Dollars: Estimation columns had different currencies. Hence, using Python language, Estimated_price_to_inDollars and Estimated_price_from_inDollars were converted to US Dollars.

3) Dropping of rows: Year is an essential factor in predicting the art price. Thus, during the extraction of data, there were a few rows which did not have a year in which the painting was made and was dropped.

4) Replacing empty cells: All the null values in the dataset were replaced with zero.

5) Dropping of columns: Less salient and highly correlated columns were removed.

6) Normalizing the columns: As columns had a different range of values, to gain efficient results, all the columns were normalized using Standard Z-score normalization.

7) Replacing artist's name with numerical value: As the model accepts only numerical values, all the nine artists' names were converted to numerical values ranging from 1 to 9.

8) Splitting of dataset: The whole dataset was divided into independent and dependent variables. To train the model, the dataset was divided into 80% training and 20% testing.

V. Exploratory Data Analysis

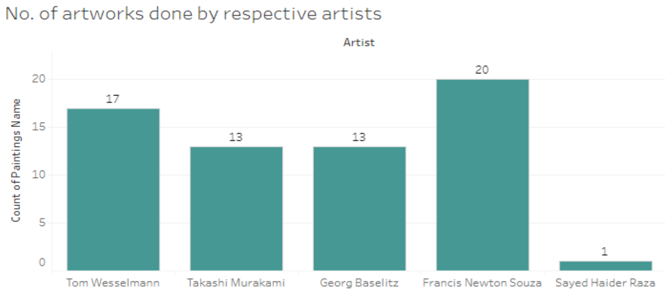
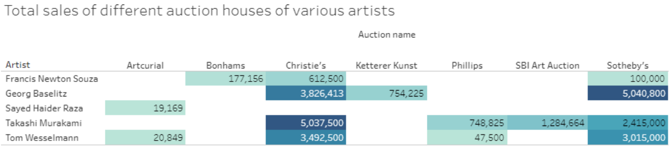


Fig. 1. Number of Artworks done by respective artist

Inference:

- In the year 2017, Georg Baselitz was the only artist with eight paintings.
- In the year 2018, Francis Newton Souza had the maximum number of paintings(20), and Sayed Haider Raza had the minimum(1).
- In the year 2019 & 2020, Tom Wesselmann had the highest number of paintings.

- Overall Tom Wessselmann and Gerhard Richter had the highest number of paintings, and Salman Toor and Sayed Haider Raza marked the least.



Total sales of different auction houses of various artists

Auction name

Artist	Artcurial	Bonhams	Christie's	Ketterer Kunst	Phillips	SBI Art Auction	Sotheby's
Francis Newton Souza		177,156	612,500				100,000
Georg Baselitz			3,826,413	754,225			5,040,800
Sayed Haider Raza	19,169						
Takashi Murakami			5,037,500		748,825	1,284,664	2,415,000
Tom Wesselmann	20,849		3,492,500		47,500		3,015,000

Fig. 2. Total Sales of every Auction along with artist name

Inference:

- In the year 2017, Christie's Auction house had the maximum sale($432,500).
- Sotheby's($5,040,800) and Christie's($5,037,500) auction house marked the highest sales in the year 2018.
- Sotheby's auction house was perpetual with the highest sales in the year 2019, 2020 & 2021.

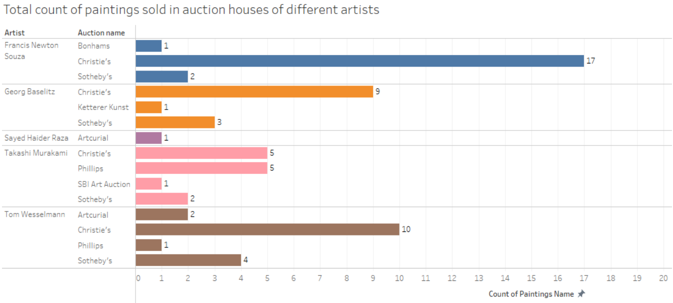


Fig. 3. Total count of paintings that each sold along with the artist.

Inference:

- Dorotheum auction house sold the most paintings(3), but Christie's had the highest sale in the year 2017.
- Christie's auction house sold 17 paintings of the artist, Francis Newton Souza in 2018.
- In the year 2019, Philips auction house sold the maximum number of paintings (12) of the artist, Takashi Murakami.
- In the year 2020, the artist Tom Wesselmann's artwork was sold maximum in Sotheby's auction house.
- Last year, Sotheby's auction house sold the highest number of paintings (23) of the artist Gerhard Richter.

VI. Methodology

A. Multiple Linear Regression

A linear regression model seeks to develop a relationship between many features (independent variables) and a continuous target variable (dependent variable). Finding the line that best fits the data is the primary goal. The line that has the lowest potential total prediction error across all data points is considered to be the best fit line. Error is the distance between the point to the regression line.

$Y = \theta_o + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$

B. Polynomial Regression

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial.
It is also called the special case of Multiple Linear Regression in machine learning as we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression. It is a linear model with some modifications to increase accuracy. The data-set used in Polynomial regression for training is of non-linear nature. It uses a linear regression model to fit the complicated and non-linear functions and data-sets. Hence, in Polynomial regression, the original features are converted into Polynomial features of the required degree (2,3,..,n) and then modelled using a linear model.

C. Random Forest Regressor

With the use of several decision trees and a method known as Bootstrap and Aggregation, also referred to as bagging, Random Forest is an ensemble methodology capable of handling both regression and classification tasks. This method's fundamental principle is to integrate several decision trees to get the final result rather than depending solely on one decision tree. [1]

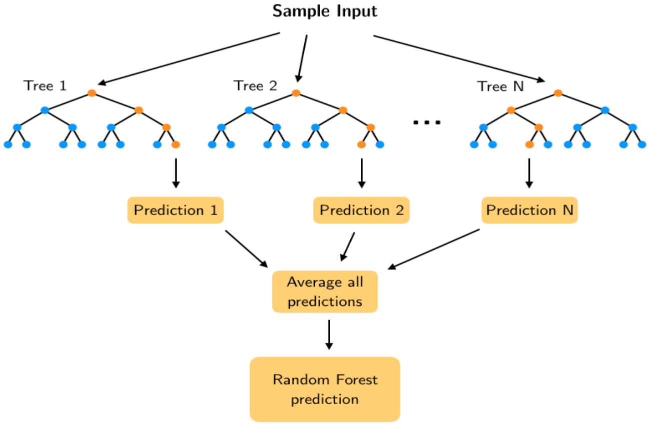


Fig. 4. Random Forest Regression

VII. Evaluation Metrics for Regression model

1) Mean Absolute Error: The MAE score is measured as the average of the absolute difference between actual and predicted values

$$\text{MAE}= (\frac{1}{n})\sum_{i=1}^{n} |y_i - x_i|$$

[1] Image Source

2) Mean Square Error: The MSE score is measured as the average squared difference between observed and predicted values.

$$\text{MSE}= (\frac{1}{n}) \sum_{i=1}^{n} \left|(y_i - x_i)^2\right|$$

3) Root Mean Squared Error: The RMSE score is measured as the root of the average squared difference between observed and predicted values.

$$\text{RMSE} = \sqrt{(\frac{1}{n}) \sum_{i=1}^{n} (y_i - x_i)^2}$$

VIII. Tools and Techniques

1) Exploratory Data Analysis: After extracting the data, we performed a few analyses to understand it better. We used Tableau. This software helps people be data-driven and is one of the best options available regarding visuals and dashboards[2].

2) Price Prediction Model: Using Google Collab, the data was pre-processed, split into training and testing, and regression models were trained. Further, evaluation metrics were used to evaluate our model. Libraries such as Pandas, NumPy, and sklearn were used.

IX. Results

This proposed model uses supervised machine learning regression approaches like Multiple Linear Regression, Polynomial Regression, and Random Forest Regressor. After performing cross validation and comparing all models based on MAE, MSE and RMSE, Multiple Linear Regression has the lowest error.

Models	**Mean Absolute Error**	**Mean Squared Error**	**Root Mean Squared Error**
Multiple Linear Regression	0.11	0.19	0.44
Polynomial Regression	0.14	0.20	0.45
Random Forest Regressor	0.16	0.41	0.64

X. Conclusion and Future Work

In conclusion, Multiple Linear Regression gives the best results compared to all the models. This project can be further improved with more rows and columns to feed the trained model. Also, websites apart from ARTSY can be used to extract data. Further, we can also use images of the art piece to predict the price using deep learning models such as CNN and LSTM.

[2] Tableau Dashboard Link

References

[1] V. Singrodia, A. Mitra and S. Paul, ”A Review on Web Scrapping and its Applications,” 2019 International Conference on Computer Communication and Informatics (ICCCI), 2019, pp. 1-6, doi: 10.1109/ICCCI.2019.8821809./

[2] K. U. Manjari, S. Rousha, D. Sumanth and J. Sirisha Devi, ”Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm,” 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), 2020, pp. 648-652, doi: 10.1109/ICOEI48184.2020.9142938.

[3] Chen Liu, ”Prediction and Analysis of Artwork Price Based on Deep Neural Network”, Scientific Programming, vol. 2022, Article ID 7133910, 10 pages, 2022. https://doi.org/10.1155/2022/7133910

[4] S. Yadav and S. Shukla, ”Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification,” 2016 IEEE 6th International Conference on Advanced Computing (IACC), 2016, pp. 78-83, doi: 10.1109/IACC.2016.25.

[5] M. H. Rahman, S. I. Nahid, I. H. Al Fahad, F. M. Nahid and M. M. Khan, ”Price Prediction Using LSTM Based Machine Learning Models,” 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2021, pp. 0453-0459, doi: 10.1109/IEMCON53756.2021.9623120.

[6] M. S. Acharya, A. Armaan and A. S. Antony, ”A Comparison of Regression Models for Prediction of Graduate Admissions,” 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-5, doi: 10.1109/ICCIDS.2019.8862140.