Data Science, Art analytics and Machine Learning Art

Ruyue Zhang, M.S.

Nima Zahadat, Ph.D.

The George Washington University

**ABSTRACT**

On 25 October 2018 at Christie’s New York, a portrait called Edmond Belamy, sold for $432,500. It is the first artwork generated by deep learning to sell in the auction house and it sold 45 times higher than its primary estimation. This portrait was created by the machine learning neural network, Generative Adversarial Network. For long time, art is considered as symbol full of conception, emotion, and abstraction. Most of the time, art analysis emphasized on the explanation of representation within the art piece. While data science offers new way to research and analyze artist, artwork in a quantitative way. The purpose of this project is to demonstrate data science and machine learning techniques for analyze art market and artwork, art auction prediction, artwork style transfer and artwork generation. The analysis visualized the trend of the art market and the color pattern in terms of different artists ‘work. The art auction prediction models adapt linear regression to forecast its hammer price and decision tree to find the importance of features concerning whether the artwork can be sold or not in an auction impacting. Image style transfer utilized pertained convolution model to update the pixel combing the style and the content to form new images. Deep convolutional GAN is experimented in the project to generate cart images between romanticism, impressionism, and realism. Finally, suggestions and improvement for further research and were provided.

*Keywords:* AI art, art analytics, machine learning, Decision Tree, Linear Regression, Kmeans, image style transfer, GAN, image generation

**INTRODUCTION**

On 25 October 2018 at Christie’s New York, a portrait called Edmond Belamy sold for $432,500. It is the first artwork generated by deep learning neural network to sell in the auction house and it sold 45 times higher than its primary estimation. This portrait was created by Obvious, a Paris-based collective of artists, Gauthier Vernier, Pierre Fautrel and Hugo Caselles-Dupré. They are interested in Machine Learning research and AI for Art. They used Generative Adversarial Network to enable a computer to create portraits of the 18th century. The generated outcomes were made into a collection of 11 realistic portraits. The Obvious even titled the 11artworks as a fictional family, The Belamy Family. A signature was given at the right lower side, which is the loss function formula of GAN. It signals that GANs begins to receive attention from the art community.

Before this, Robbie Barrat, uploaded his Landscape and Nude portrait GAN Lua code to GitHub at his 17. He fed the network 15000 pieces of landscape paintings and nude portraits to train it over two weeks. When Obvious created the Belamy Family, they borrowed Robbie Barrat code.

After Edmond Belamy was sold, Mario Klingemann’s GAN work is going to sale at Sotheby’s contemporary art auction house. He is a German artist and Google Arts and Culture resident known for his work involving neural networks, code, and algorithms. He is considered a pioneer in the use of machine learning arts His works examine creativity, culture, and perception through machine learning and artificial intelligence, and have appeared at the Ars Electronica Festival, the Museum of Modern Art New York, the Metropolitan Museum of Art New York, the Photographers’ Gallery London, the Centre Pompidou Paris, and the British Library. In 2018, he won the Lumen Prize which recognizes works of art made using technology.

In terms of data science for art, date science offers new perspective and techniques to understand and analyze art. In the past, art analysis focusses on about the element, contextual and historical in a descriptive way. Taking advantage of Data Science, it’s faster and easier to qualitative data to visualize connections, relationships, and dynamics between a single artist’s work or among numbers of artist, or even entire work series of an art movement. Meanwhile, more and more art collection is shared online through museum website, or website like WIKI art and Google Arts & Culture. With large art datasets public, it’s time for art to meet big data.

**LITERATURE REVIEW**

In 1967, Leslie Mezei mentioned in his article *Computers and the Visual Arts* that While the monumental tasks of museum cataloging and art history are being considered (CHum, [1966-67]), a start has yet to be made in the analysis of art (Leslie Mezei, 1967). Expensive scanning equipment, large memories and complex software are required. But at that time, although scientists are processing moon photographs, bubble- chamber tracks, chromosome pictures, and aerial photographs for "target detection," little parallel effort has been directed to works of art (Leslie Mezei, 1967). As the computer and its graphic capabilities become easily available to these groups another exciting spurt will develop toward the "Second Renaissance” ((Leslie Mezei, 1967).

After years of development for computational power, research on artificial intelligence art appeared. Developments in Al already allow artists to go beyond the possibilities of the past, for example by making artworks that respond differently for different viewers, as in the case of a Computer Art kinetic sculpture that, upon being informed about a viewer’s age, performs motions the artist-programmer chose to be appropriate for that age(Stephen,1983). The heart of the simulated personality provided by an example of A1 Computer Art is the conceptual information inherent in its program, and it can exist and unfold only through a process of collaborative interaction between viewer and machine(Stephen,1983). Developments in A1 might permit artists to make artworks with human-like sensibilities, those that interact with viewers in ways considered intelligent and those that learn from experience(Stephen,1983).

Indeed, now machine learning is one of the most cutting-edge advancements in computer science and with the rise of deep neural network, the algorithm for image style transfer and image generation established.

A Neural Algorithm of Artistic Style that can separate and recombine the image content and style of natural images was introduced in 2016. The generic feature representations learned by high-performing Convolutional Neural Networks can be used to independently process and manipulate the content and the style of natural images (Leon A,2016). The algorithm allows to produce new images of high perceptual quality that combine the content of an arbitrary photograph with the appearance of numerous well-known artworks (Leon A, 2016). The results provide new insights into the deep image representations learned by Convolutional Neural Networks and demonstrate their potential for high level image synthesis and manipulation (Leon A,2016).

Generative Adversarial Networks (GANs) are generative models created in 2014 by Ian Goodfellow. They simultaneously trained two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake (Ian Goodfellow, 2014). Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples (Ian Goodfellow, 2014). Since, research papers on using different structure and loss function of GANs for Art showed, for instance, CycleGAN, BigGAN. The creative potential of this algorithm is still be studied.

**RESEARCH METHODOLOGY**

Image data were scraped by python 3.5, and the package Beautifulsoup. Auction data used Python3.5, Selenium and chrome driver to scrape in an automated measure. Data cleaning and data saving primarily made by Pandas and Numpy. The cleaning process includes excluding null value, changing data type, regulating data format. Visualization applied Matplotlib and Plotly to exhibit the art market trend and color pattern for artists. The four models, Linear regression, Decision Tree, image style transfer, image generation from GAN were done by Sklearn, Tensorflow and Keras. The Tensorflow and Keras models were trained on the google cloud platform with Tesla K80 GPU. At last, analysis and conclusions were drawn based on the results of all progression in the project.

**DATA**

The entire image data was from WIKIART website and they are all in public domain. There is no illegal problem. Auction data was from Askart.com and it’s available by its subscribers. George Washington University has subscribed it. The data was acquired under University’s IP address and will only be used in this project. Image data consists of painting by Vincent Van Gogh, Claude Monet, J.M.W.Turner, Edward hoper . Total number for image data is 4619. They are used in the color pattern visualization, image style transfer and image generation with DCGAN. The auction data contains auction records for Vincent Van Gogh, Claude Monet, Edward hopper, Aguste Renoir, Camille Pissarro around 3500. The auction data for every artist comprise artwork title, hammer price, high estimate price, low estimate price, signature, size, create date, medium, auction house, auction lot and auction date. Size was split into width and height. And the sold status was added to auction data indicating the artwork is sold or not. All auction data were saved into csv. Art market data were downloaded as excel file from Statista. They were visualized to manifest the art market trend. The art market data describes the percentage of the worldwide auction market in 2017 by country, market share of fine art auction revenue worldwide in 2018 by country, sales at public art auctions in the us from 2010 to 2017, volume of transactions in the art market worldwide from 2007 to 2017, percentage of art auction revenue worldwide from 2016 to 2017 by city and art type, number of fine art lots sold at the leading auction houses worldwide 2018, leading auction houses worldwide in 2018 by fine art revenue, impressionist and post-impressionists leading artists by auction turnover 2017, and reasons for collecting art among art collectors in the US as of February 2018.

**DATA ANALYSIS**

For auction data, at first, they were saved into Pandas data frame after scraping. When data cleaning process was done, they were saved into csv file. Every column in the auction data is rather irregular due to marks like exclamation mark, question mark, colon. These marks were immediately replaced after saved into Pandas data frame. NaN values were initially replaced by 0 instead of dropping. Because some artworks did not get sold and their hammer price is NaN. In order to predict if an artwork will be sold or not, the hammer price with NaN should not be dropped. While before fitting into the linear regression model, all rows with 0 in high estimate price column and low estimate price column were dropped. Data in columns like hammer price, high estimate price, low estimate price were changed into integer. Then all string features were encoded into numbers by Sklearn LabelEncoder, such as medium of the painting, auction house, title and artist name. As mentioned above, size column was split into width and height in centimeters. It was dropped after the width column and the height column created. Sold status was added to auction data and encoded into 0 and 1 to imply whether the artwork is sold or not. And this made it qualified for the decision tree model. The node of the decision tree was visualized to signify the feature importance among all the features according to information gain. The Sklearn package provided support for all modeling requirements, to include splitting and scaling the data, fitting the training data, scoring the model and predicting the outcomes of test data.

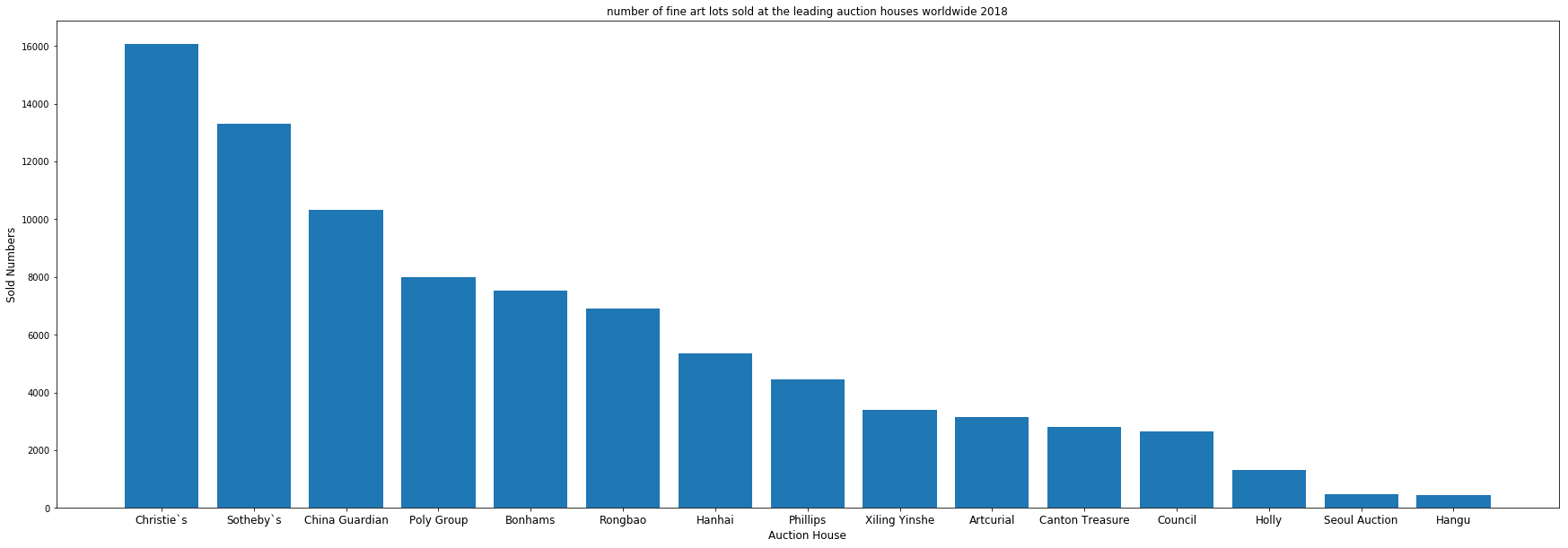
Visualization for the art market is made up of maps, Line chart, bar chat, scatter plot. The original art market data files are in excel format. They were loaded with Pandas and visualized by Matplotlib and Plotly. The image data are mainly prepared for the color cluster K-means model, the image transfer style model and the deep convolution generative adversarial network. When feeding to the model, all images were processed into array with float between -1 and 1 then vectors with shapes following numbers, image width, image height and image channel. Finally, the data would be transferred back from 0 to 1 when the image needed to show in RGB model. The code is publicly available at <https://github.com/samroberts58/Phishing-Website-Metadata-Analysis>.

**KEY FINDINGS**

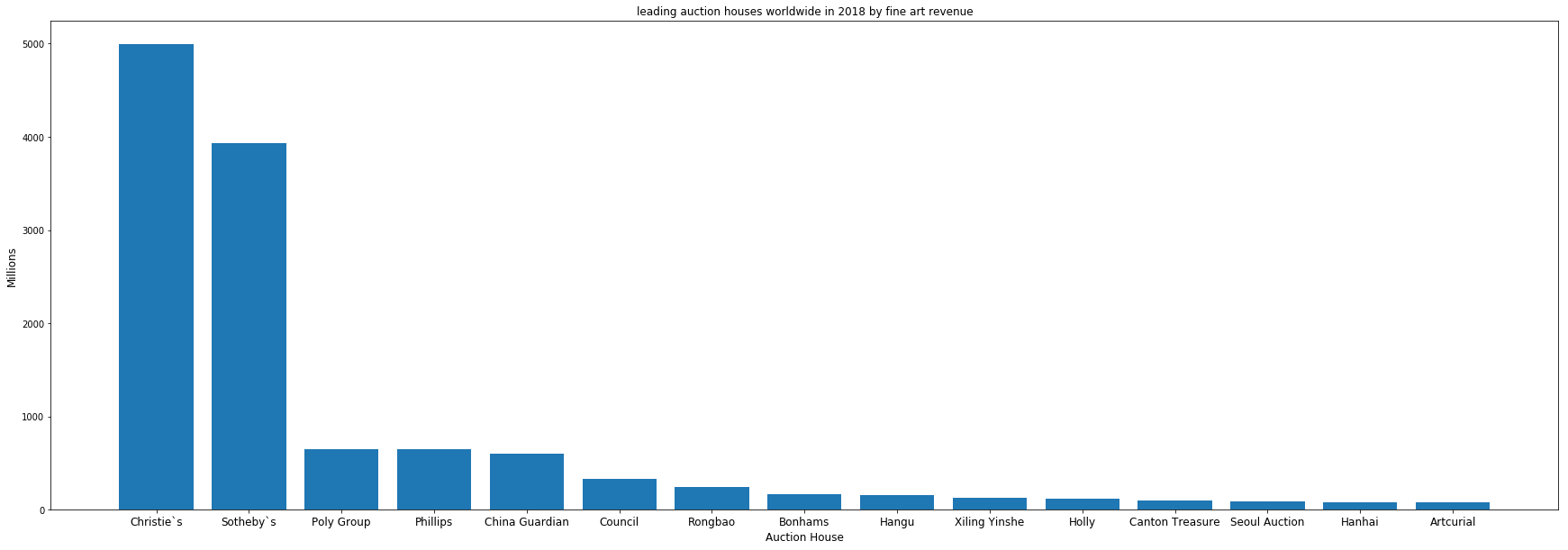
* **Art market and Art auction hammer price prediction model**

The auction data and art market data are associated with each other. When exploring the whole auction data, Sotheby's New York, Christie's New York, Rockefeller Center, Sotheby's London, Christie's London,

Christie's Paris are the top auction house holding the most auction. In the visualization for number of fine art lots sold at the leading auction houses worldwide 2018 and leading auction houses worldwide in 2018 by fine art revenue, Christie are in the first place followed by Sotheby.

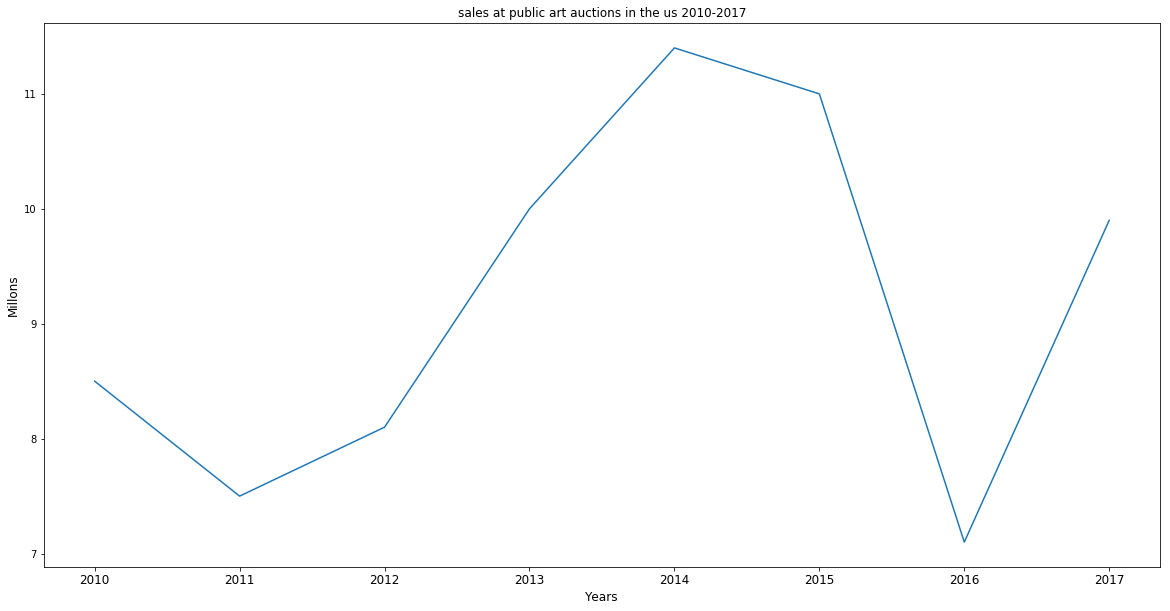


*Figure 1*. Number of fine art lots sold at the leading auction houses worldwide 2018.

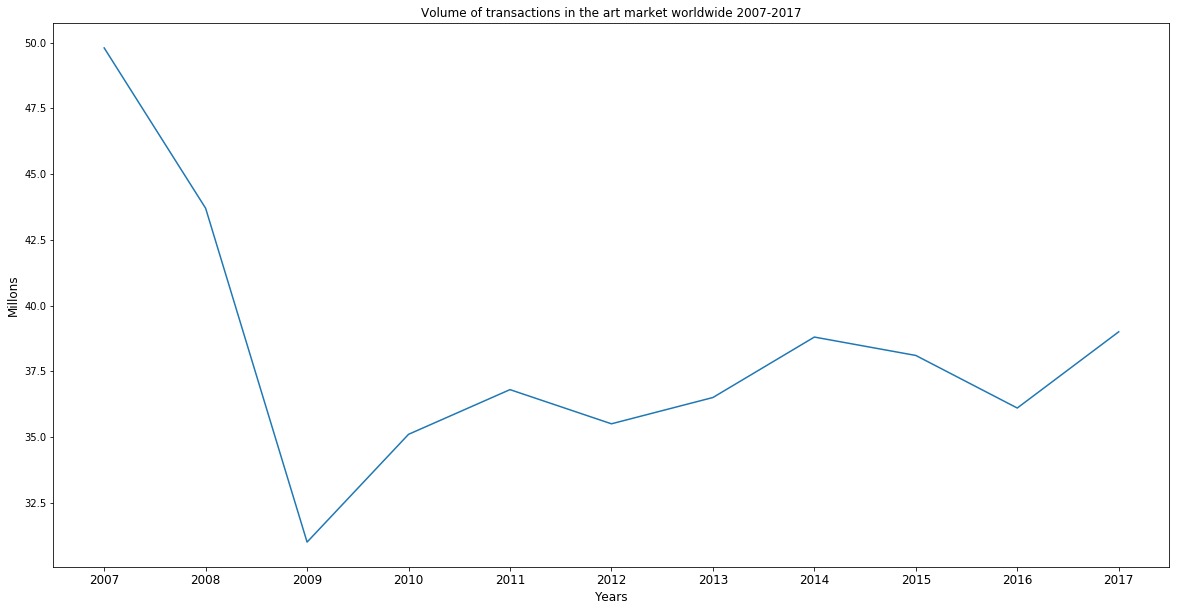


*Figure 2*. leading auction houses worldwide in 2018 by fine art revenue.

At the same time, the US is the country makes up the largest percentage of the worldwide auction market in 2017 and market share of fine art auction revenue in 2018. Meanwhile, New York shares the largest percentage of art auction revenue worldwide from 2016 to 2017 in contemporary art and prior periods. Both the Sales at public art auctions in the us from 2010 to 2017 and the volume of transactions in the art market worldwide from 2007 to 2017 had reached a low peak at 2016 and raised again after 2016.

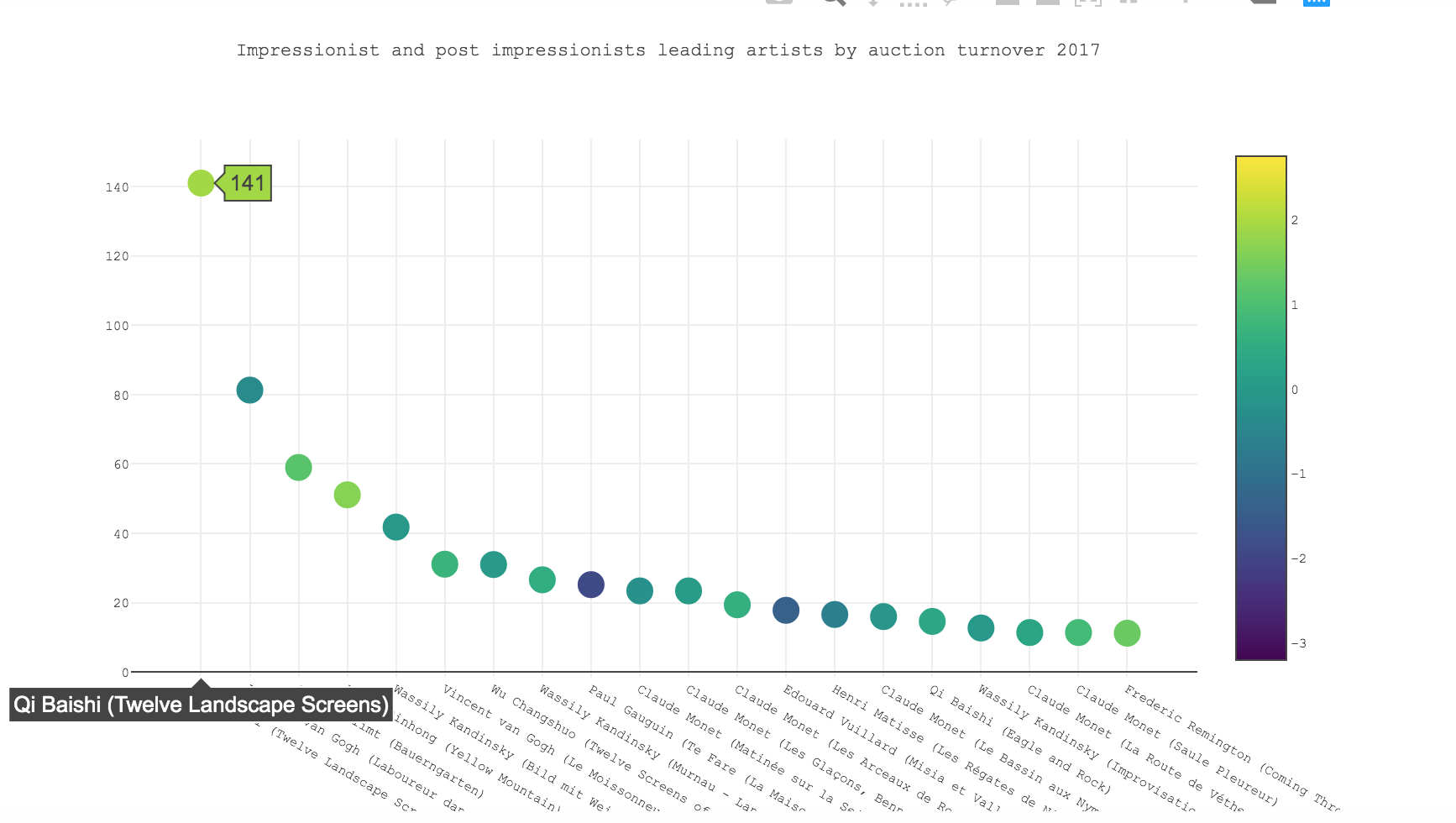


*Figure 3*. Sales at public art auctions in the us from 2010 to 2017.



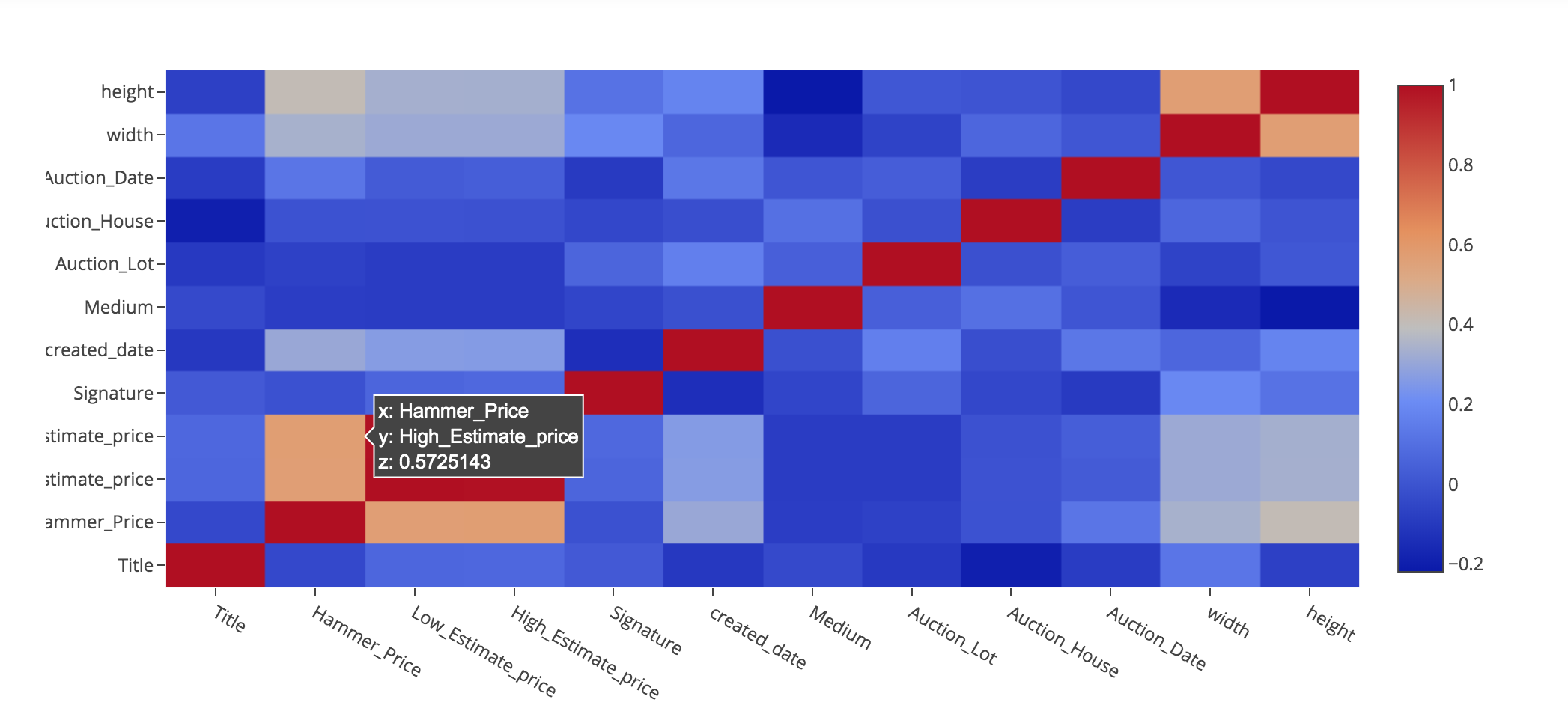
*Figure 4*. Transactions in the art market worldwide from 2007 to 2017.

The image data for this project selected from many artists in impressionism movement. The visualization of impressionist and post-impressionists leading artists by auction turnover in 2017 unexpectedly proved that the Chinese Painter Qi Bashi ranked the highest place with Monet in the second.

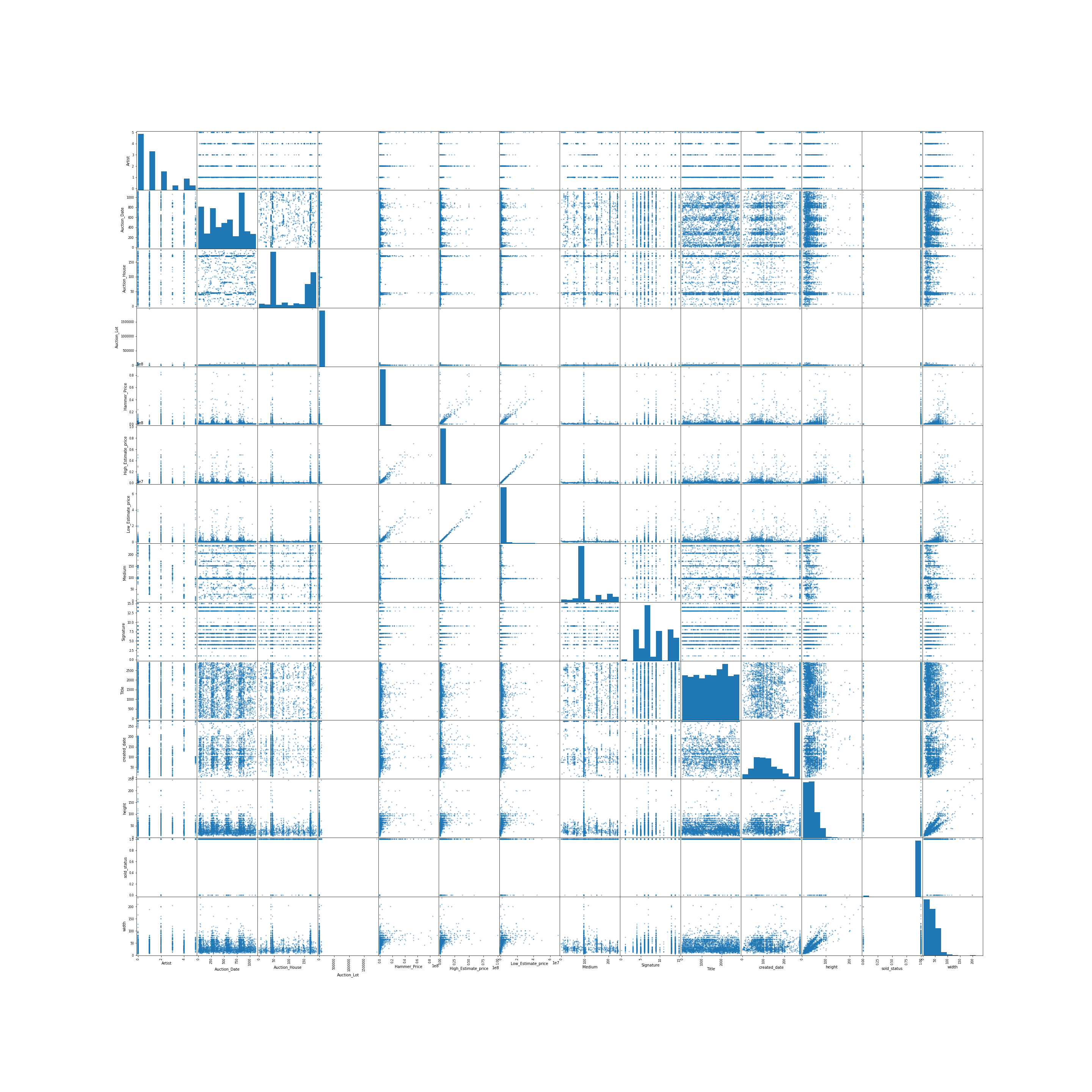


*Figure 5*. impressionist and post-impressionists leading artists by auction turnover in 2017.

Linear regression models were set for the auction data to predict the hammer price for artworks in an auction. First small size of train data was fit into the model containing three artists’ work, Van Gogh, Hopper and Monet. The measure for the model score in Sklearn is default to R squared. The R squared score for the first model is as low as 0.48 and the mean squared error is huge. which means the model did not fit the data quite well. After raising the size of data to 3520, the second linear model’ R squared was at 0.44, which is even lower than the first model. While in the process of adding the data, the zero value in the column high estimate price increased a lot. Examining the correlation map and the scatter matrix, it’s clear that high and low estimate price are relatively strong correlated with hammer price comparing to other features.

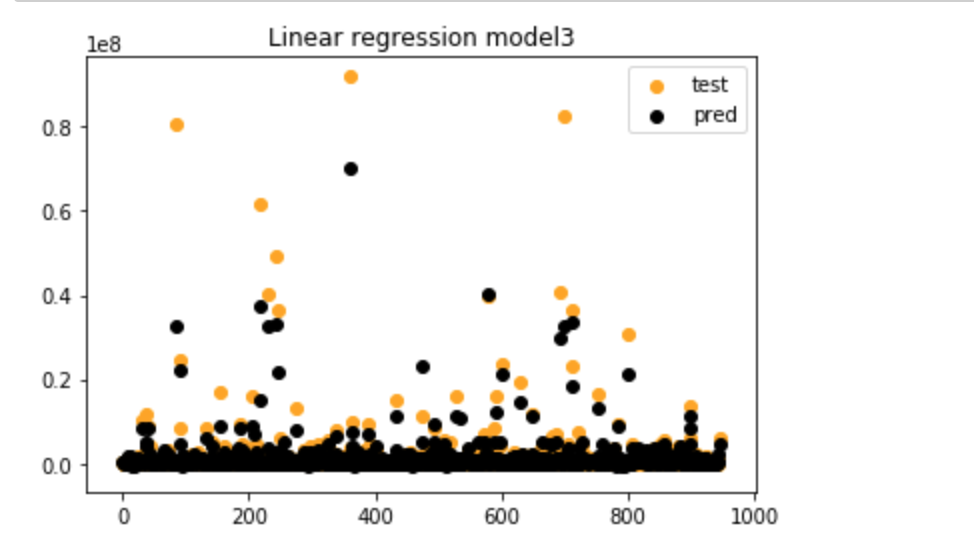
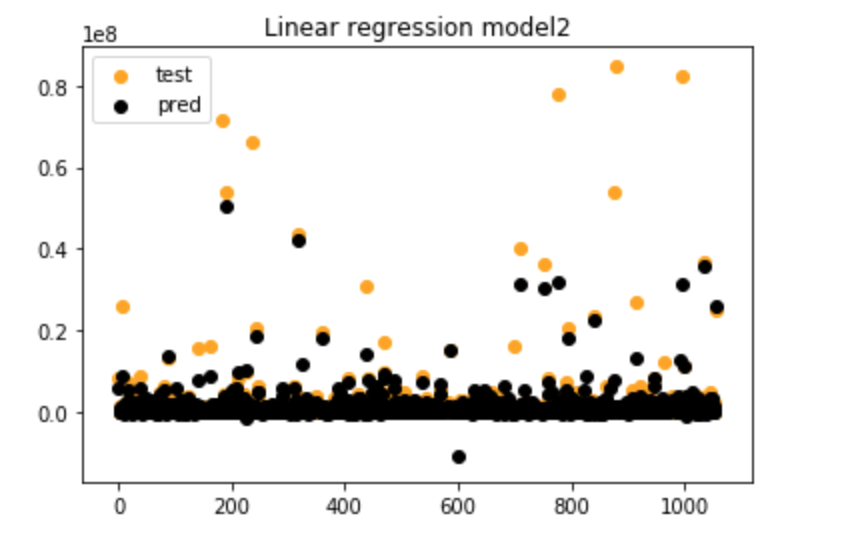


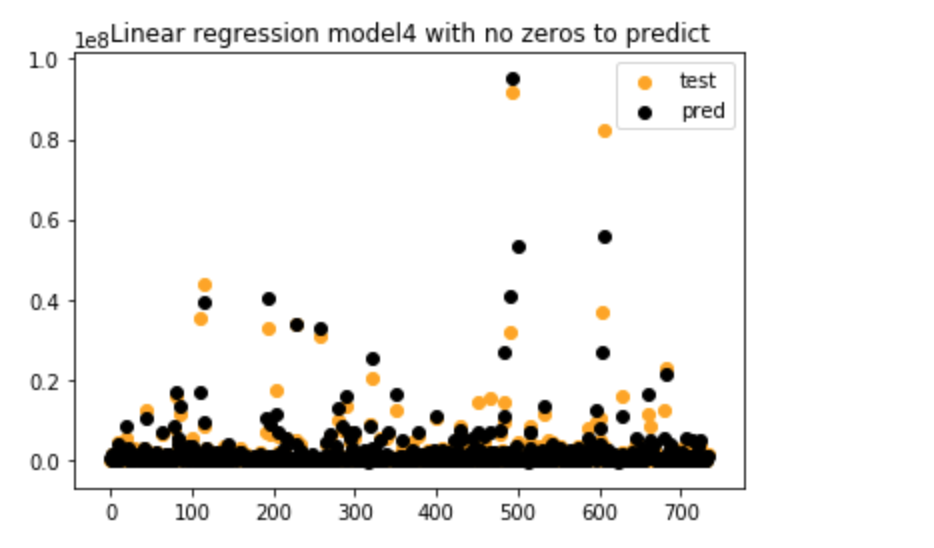
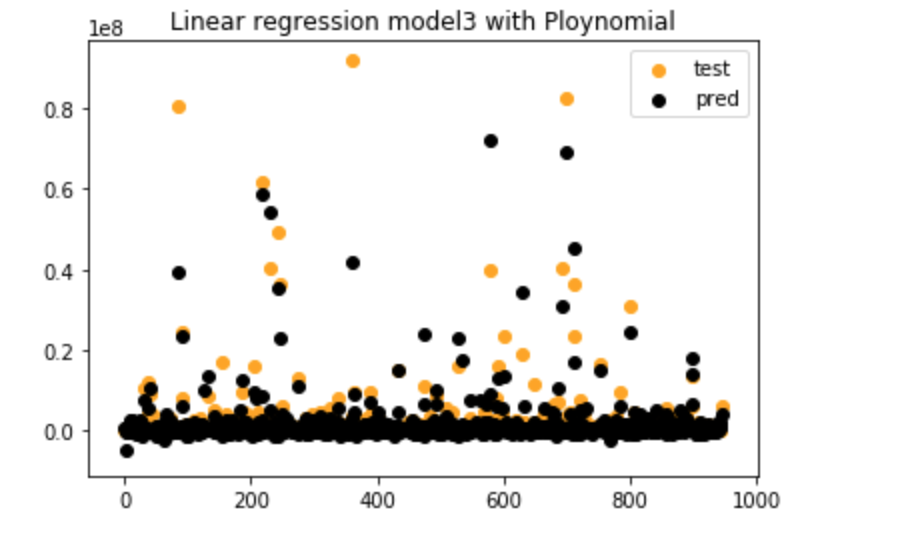
*Figure 6*. Correlation heat map for Van Gogh’s auction data.



*Figure 7*. Scatter matrix for auction data with 3520 records.

The third model is based on the data dropped all rows covering zero in high estimate price. The R squared for the model improved to 0.80. However, the mean squared error kept huge. Seen from the following figure, there are some big outliers are hard to predict for the two models.



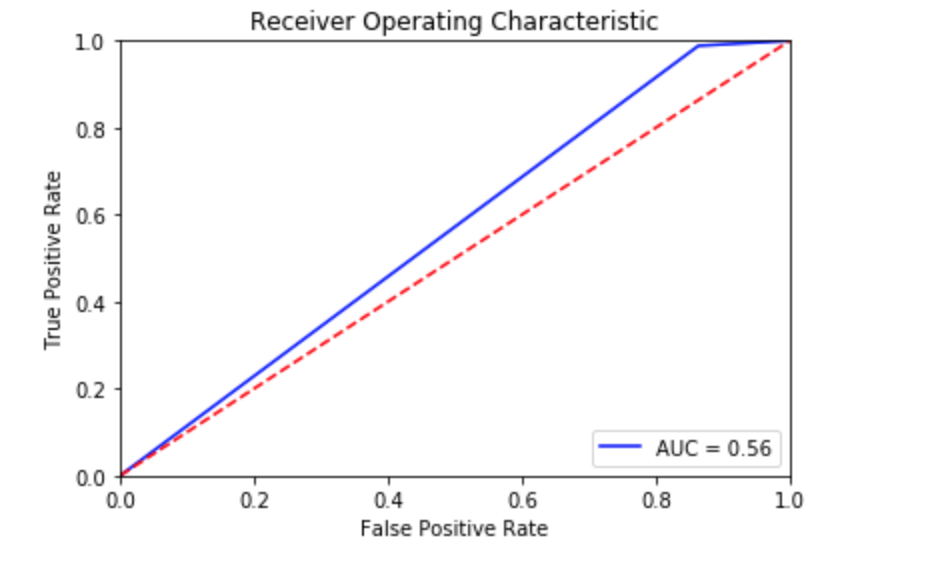


*Figure 8*. The comparison between true hammer price and predict hammer price of the four models.

In order to match the outliers, the fourth model adopted polynomial regression with degree 2. The R squared (0.77) did not improve while the visualization indicates that it did match some outliers. Model 4

Based on the input data removed rows containing zero in hammer price. Zero value in hammer price represents this artwork not sold. Removing them is to predict a piece's hammer price on the premise that it is sold. Although the R squared rise to 0.83, but considering the intention for the model is to offer estimate hammer price for an art piece awaiting sale. The fourth model is not a practical model. The ridge and lasso regularization for the coefficients were also tried on the second model. The revised model did not outperform the original model comparing mean square error.

The decision tree model to predict whether the artwork can be sold at the auction did reach high accuracy around 0.97. While checking with confusion matrix, because this dataset is unbalanced. With not sold art pieces only account for 2%. The possibility is that the model can keep guess the art sold all time to get high accuracy. This possibility was confirmed with the Area Under the Curve.



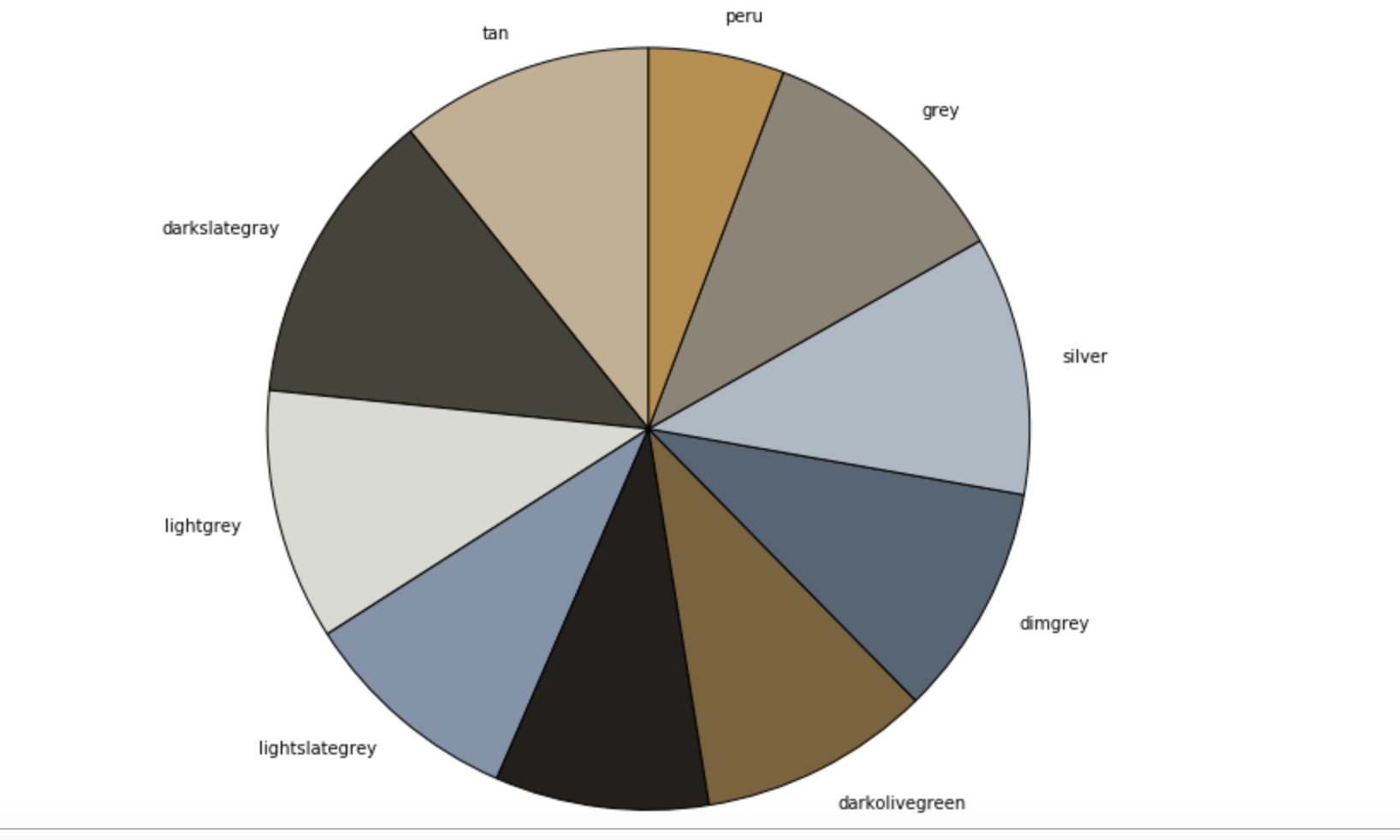
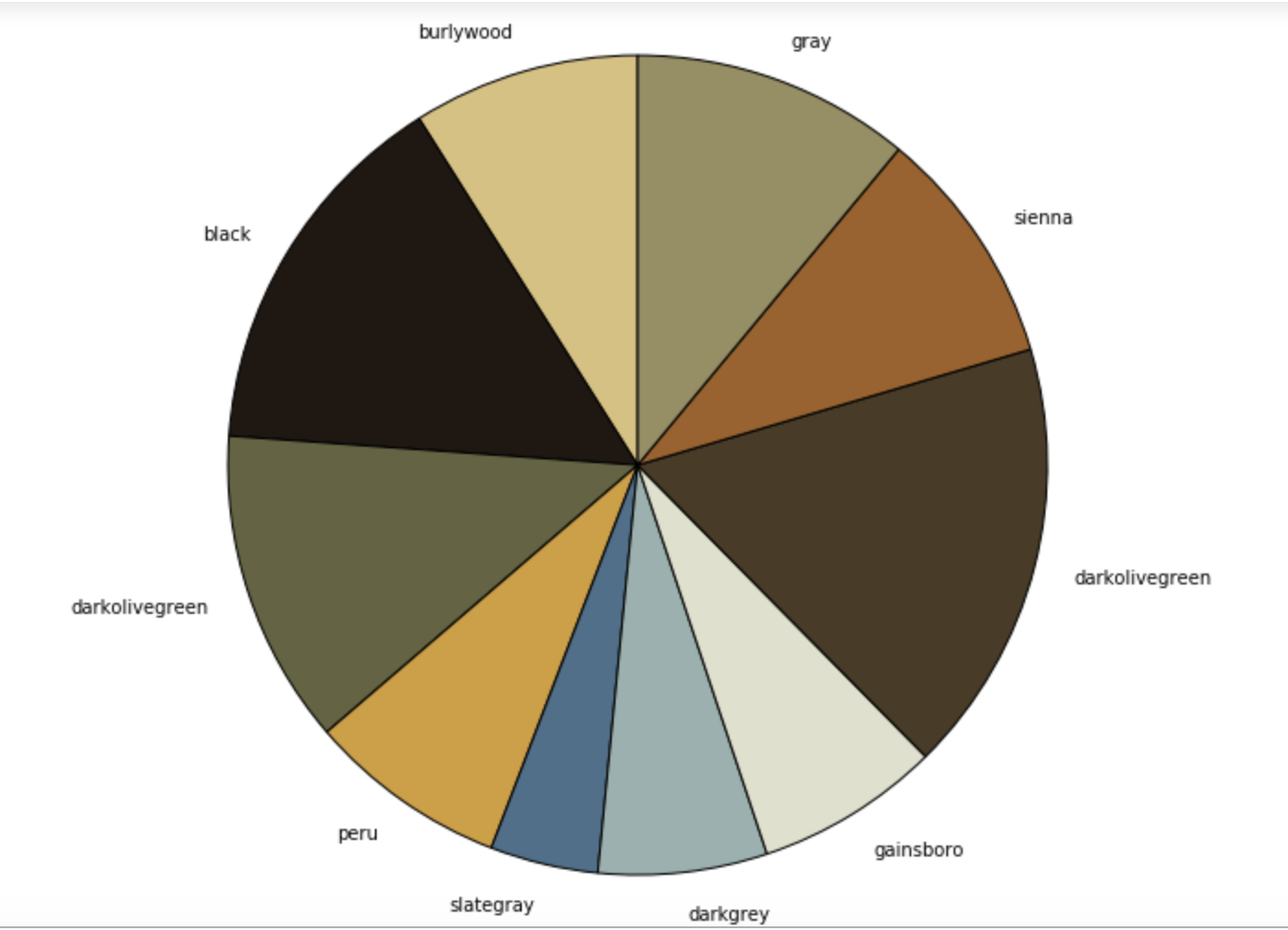
*Figure 9*. The Receiver Operating Characteristic and Area Under the Curve.

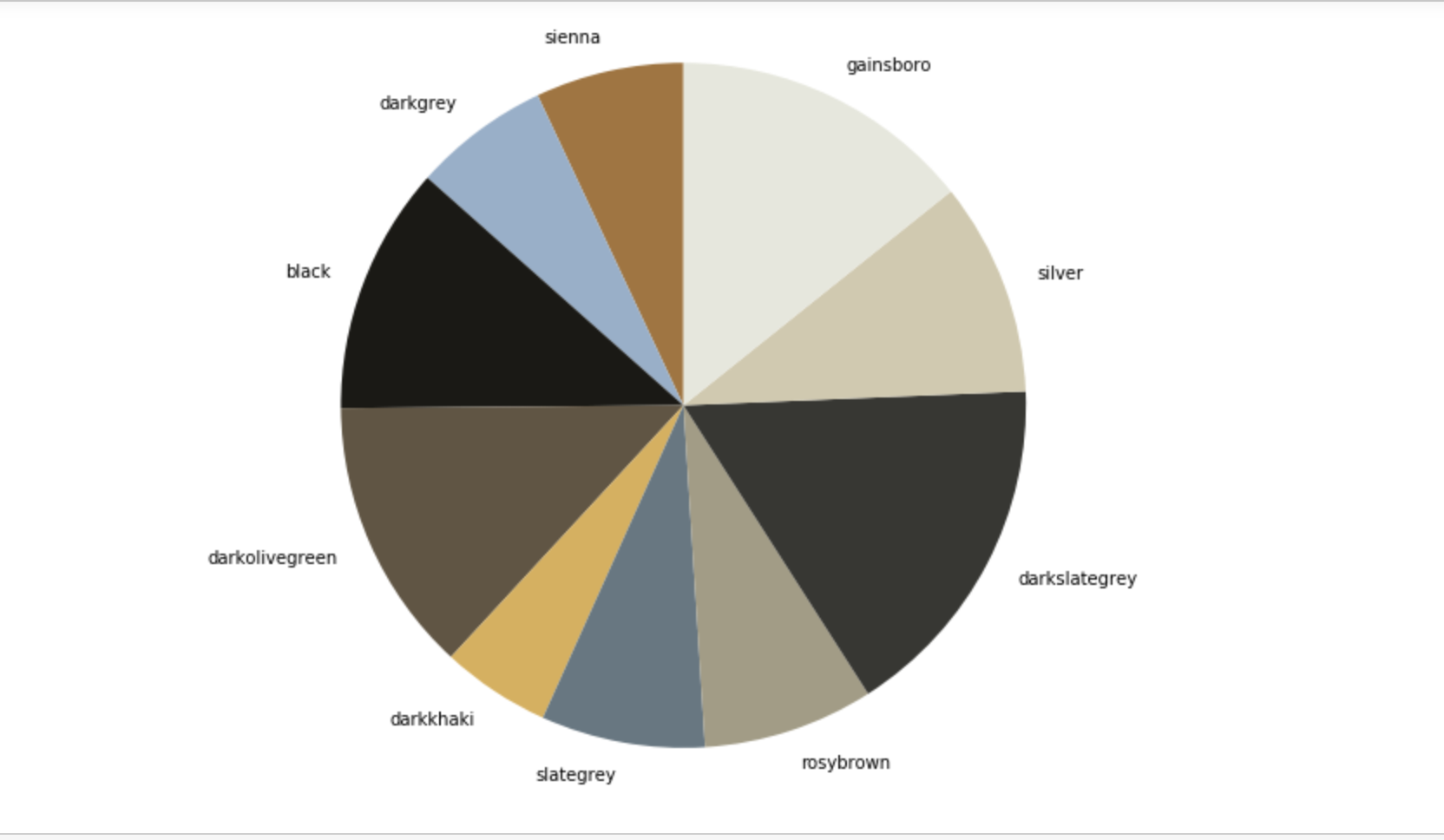
The AUC is 0.56. It means that the capacity of the model to separate class is not good. The biggest problem lies in the dataset. Besides unbalance in the sold status, numbers of painting to be sold is also disproportionate. The number of Pissaro’s work sold at the auction is almost 10 times of Van Gogh’s work. This impacted the model heavily as the artist feature is the root node in the decision tree, which means it is the best predictor for the model. Therefore, to enhance the dataset is the most essential point for building a better predict auction price model.

##!!!and reasons for collecting art among art collectors in the US as of February 2018.

* **Color cluster by K-means**

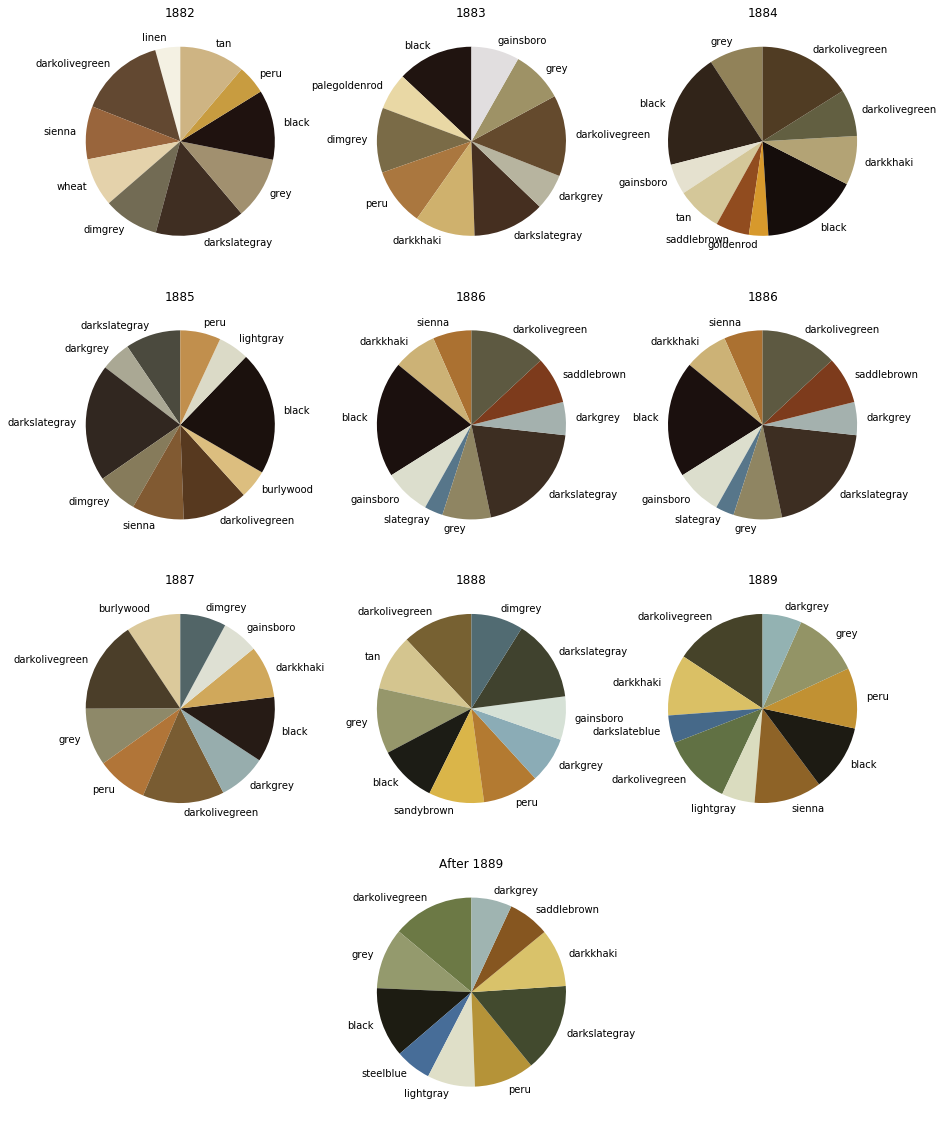
Monet, Van Gogh and Hopper’s images were chosen for the analysis about the color pattern. By converting all images from one artist into a vector then fit in the K-means model to find 10 clusters of color.





*Figure 10*. Color clusters for Van Gogh, Monet and Hopper.

Above are the top 10 color clusters for Van Gogh, Monet and Hopper. Surprisingly, those color pattern are similar to each other, especially the cluster for Hopper and Monet (the last two). Hopper’s color cluster is a lighter version of Monet. Despite that Hopper is a symbol of American realism, and Monet is the one of the pioneers in impressionism. The impression Hopper’s painting made on people is the emotion isolation and lioness. Yet Hopper did greatly admire the impressionists of Europe.The same as Monet, light is a very crucial element in his painting. The color pattern further proved this.



*Figure 11*. Color clusters from different years for Van Gogh.

As for Van Gogh, his color pattern changed from classical Dutch painting to impressionism to post impressionism. Each stage the color became more vivid than the previous one although he was even suffered from depression. More yellow and blue was displayed in his later years. It was said because Van Gogh was infected with yellow version which is formally called “Xanthopsia”. It overrides yellow bias in vision that can provoke the reddish-brown filter of nuclear sclerosis. Another saying is simply after Van Gogh moved to Paris, then Ares, he enjoyed more sunny days with wheat fields and yellow sun flowers

* **Image style transfer for Monet and Turner**

Both J.M.W.Turner and Monet was famous for their use of light. In this image style transfer model, Monet’s impression sunrise and Turner’s the burning of the houses of parliament were be combined. Regarding the image style transfer, it uses the pre-trained convolution neural network to update the pixel of the combined image rather than the conventional convolution neural network that keeps updating model’s weight via backpropagation. During the training for image style transfer, weight remained the same.

Knowing that we can distinguish layers that are responsible for the style (basic shapes, colors etc.) and the ones responsible for the content (image-specific features), we can separate the layers to independently work on the content and style.

Then we set our task as an optimization problem where we are going to minimize:

content loss (distance between the input and output images - we strive to preserve the content)

style loss (distance between the style and output images - we strive to apply a new style)

total variation loss (regularization - spatial smoothness to denoise the output image)

et our gradients and optimize with the L-BFGS algorithm

As it can be seen, in image generated using style weight 3200 painter’s brushstrokes are much more prominent. Original content (colors etc) is much less prominent.

In Style transfer learning, we are going to use a deterministic optimizer l-bfgs instead of Stochastic Gradient Descent or Adam because:

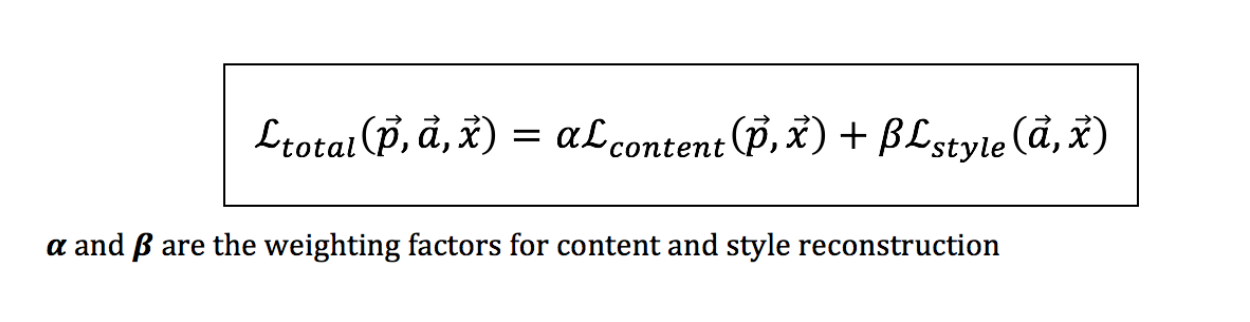
There is nothing stochastic here (we are not creating random mini-batches with different set of data), instead, the optimizer keeps getting the same single image.

l-bfgs determines which direction to go and the optimum distance to go in that direction by computing the hessian and doing a line search. This approach is expensive in stochastic problems but it is the right approach here.

l-bfgs does learn faster than Adam in style transfer.

We can change the hyperparameters to control how much we want to match the content versus how much we want to match the style.





The idea is first presented by Gatys et al. Basically, you’ll have two input images, a content image and a style image. And we wanna produce a mixed image which contains style (such as texture, colour) from style image and content from content image. In convolutional neural network, the feature maps in lower layers captures the low level features (i.e. content) and vice versa, the feature maps in higher layers captures the high level features (i.e. style). The image shown below illustrates the idea. During style reconstructions, the higher feature levels capture the style of the painting. On the other hand, during content reconstructions, the lower feature levels capture the content.

artificial system based on a Deep Neural Network that creates artistic images

**of high perceptual quality. The system uses neural representations to separate and recombine content and style of arbitrary images, providing a neural**

**algorithm for the creation of artistic images. Moreover, in light of the striking similarities between performance-optimised artificial neural networks and**

**biological vision,3–7 our work offers a path forward to an algorithmic understanding of how humans create and perceive artistic imagery.**

**Generative Adversarial Networks (GANs) are generative models created in 2014 by Ian Goodfellow, a researcher in Machine Learning. They put two algorithms in competition one with another to perform training.**

**A generator will create new images by mimicking characteristics of images from the training dataset, and try to fool a discriminator into thinking those images are “real”. The generator trains until no difference can be made by the discriminator.**

the original 2015 paper by Gatys et al. proposed a neural style transfer algorithm that does not require a new architecture at all. Instead, we can take a pre-trained network (typically on ImageNet) and define a loss function that will enable us to achieve our end goal of style transfer and then optimize over that loss function.

we start from a blank image composed of random pixel values, and we optimize a cost function by changing the pixel values of the image.

we start from a blank image composed of random pixel values, and we optimize a cost function by changing the pixel values of the image. In simple terms, we start with a blank canvas and a cost function. Then we iteratively modify each pixel so as to minimize our cost function. To put it in another way, while training neural networks we update our weights and biases, but in style transfer, we keep the weights and biases constant, and instead, update our image.

*Figure 3*. Distribution of feature importance.

**RECOMMENDATIONS**

The identification of human-determined metadata patterns for prediction could be further supported by creating and testing datasets with variables that share that main qualification of human choice. Examples of other metadata variables that could fulfill this would be themes of content chosen to attract a user’s attention, phrase analysis with natural language processing (NLP), number of JavaScript actions on the page, or information embedded within the URL. Additionally, other host and network metadata factors at various levels of the OSI model may support selection of useful variables with the intent to identify important secondary effects that human behavior influences. Another distinct approach may be to consider event-based patterns, such as socioeconomic, political, or cultural shifts that could be informative for cybersecurity monitoring and defense strategies. Reinforcement learning cycle gan better for trabs many gans data size auction generation image emotion The team collected a set of 15,000 portraits from online art encyclopedia WikiArt, spanning the 14th to the 19th century,

**CONCLUSION**

**Future market penetrate**

Phishing websites are one of the most effective and dangerous forms of cybersecurity attacks. Experienced malicious users can cast a wide net for exploitation with few resources and time invested, and with even a low response rate, can be extremely profitable. There is certainly a mix of risk factors that may make an attacker more effective and an unsuspecting user more vulnerable. Few researchers have dedicated attention towards profiling malicious users, with most studies focused towards user education and social engineering responses. This data mining project attempted to fill an informational gap and begin a conversation about the identifiable trends and human behavior patterns hidden within the metadata of phishing websites. The accuracy results of the models were low, and therefore, clearly other important features for this real-world phenomenon are missing. One positive outcome was attacker-defined metadata appeared to be more significant than the computer-generated variables. Continued trend analysis with revised variable sets has the potential to provide strong support to profiling attackers and defending against the enduring problem of phishing websites.

The potential applications of A1 in the visual arts add challenging perspectives on theoretical debates about the relationship between artists and viewers and between artistic processes and artworks and about the fundamental characteristics of aesthetic qualities of artworks.

There are a lot more examples of Art with GANs, such as the work of Michael Tyka, Samim, Alex Champanard, and several research papers on the topic…

We believe some of the work in AI and Art raises philosophical as well as societal questions. Is an algorithm capable of creativity ? This question has risen many times since computers have been created, but this time, with algorithms like GANs, it takes it a step further.

Many questions are also raised in the field of Art. In contemporary art, the artist has always been at the center of the work, and the tool as a way for him to express, and pass on emotions. Here, the tool is closer to the center of the work, even though the artist behind the algorithm remains the “real” artist. The intention and inspiration comes from the human who designed and used the algorithm. Hence the collaboration between human and machine has never been so close.

**BIOGRAPHY**

**Ruyue Zhang** is a graduate student in the Data Science Program at The George Washington University. Her interests includes bringing art with data and machine learning. She has studied Journalism in undergraduate study and thus interested of social media,history and culture. She enjoys doctor who, petting a cat, and concerts.

**Dr. Nima Zahadat** is a professor of data science, information systems security, and digital forensics. His research focus is on studying the Internet of Things, data mining, information visualization, mobile security, security policy management, and memory forensics. He has been teaching since 2001 and has developed and taught over 100 topics. Dr. Zahadat has also been consultant with the federal government agencies, the US Air Force, Navy, Marines, and the Coast Guard. He enjoys teaching, biking, reading, and writing.

**REFERENCES**

<http://time.com/5357221/obvious-artificial-intelligence-art/>

Mezei, L. (1967). Computers and the Visual Arts. Computers and the Humanities, 2(1), 41-42. Retrieved from <http://www.jstor.org.proxygw.wrlc.org/stable/30203951>

Wilson, S. (1983). Computer Art: Artificial Intelligence and the Arts. Leonardo 16(1), 15-20. The MIT Press. Retrieved April 27, 2019, from Project MUSE database.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).  [**arXiv:1406.2661**](https://arxiv.org/abs/1406.2661)

Elgammal, A.M., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. ICCC. [**arXiv:1706.07068**](https://arxiv.org/abs/1706.07068)

Alvarez-Melis, D. (2017). The Emotional GAN : Priming Adversarial Generation of Art with Emotion.

Elgammal, A.M., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. ICCC. [**arXiv:1706.07068**](https://arxiv.org/abs/1706.07068)

Gatys, L.A., Ecker, A.S., & Bethge, M. (2015). A Neural Algorithm of Artistic Style. CoRR, abs/1508.06576.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. CAN: Creative Adversarial Networks, Generating “Art" by Learning About Styles and Deviating from Style Norms. arXiv preprint arXiv:1706.07068, 2017.

Gatys, L.A., Ecker, A.S., & Bethge, M. (2015). A Neural Algorithm of Artistic Style. CoRR, abs/1508.06576.

Bullée, J. H., Montoya, L., Pieters, W., Junger, M., & Hartel, P. (2017). On the anatomy of social engineering attacks-A literature-based dissection of successful attacks. Journal of Investigative Psychology and Offender Profiling, 15(1), 20-45. doi:10.1002/jip.1482

Curtis, S. R., Rajivan, P., Jones, D. N., & Gonzalez, C. (2018). Phishing attempts among the dark triad: Patterns of attack and vulnerability. Computers in Human Behavior, 87, 174-182. doi:10.1016/j.chb.2018.05.037

Gowri, R., Gandhi, V., & Suriakala, M. (2017). An Efficient Algorithm to Identify Phishing Sites Using URL Domain Features. International Journal of Advanced Research in Computer Science,8(7), 508-510. doi:10.26483/ijarcs.v8i7.4309

Gupta, B. B., Tewari, A., Jain, A. K., & Agrawal, D. P. (2016). Fighting against phishing attacks: State of the art and future challenges. Neural Computing and Applications, 28(12), 3629-3654. doi:10.1007/s00521-016-2275-y

Gürhan-Canli, Z., & Maheswaran, D. (2000). Cultural Variations in Country of Origin Effects. Journal of Marketing Research, 37(3), 309-317. doi:10.1509/jmkr.37.3.309.18778

Jansson, K., & Solms, R. V. (2013). Phishing for phishing awareness. Behaviour & Information Technology, 32(6), 584-593. doi:10.1080/0144929x.2011.632650

Liu, W., & Zhong, S. (2017). Web malware spread modelling and optimal control strategies. Scientific Reports, 7(1). doi:10.1038/srep42308

Rajupra, H. & Diwanji, H. (2013). Enhancement of Fake Website Detection Techniques Using Feature Selection and Filtering Algorithms. International Journal of Advanced Research in Computer Science, 4(3). doi: [10.26483/ijarcs.v4i3.1551](https://doi.org/10.26483/ijarcs.v4i3.1551)

Roberts, S. (2018). samroberts58/Phishing-Website-Metadata-Analysis: Phishing Website Metadata Analysis (Version v1.0.0). Zenodo. <http://doi.org/10.5281/zenodo.2248885>

Schaik, P. V., Jeske, D., Onibokun, J., Coventry, L., Jansen, J., & Kusev, P. (2017). Risk perceptions of cyber-security and precautionary behaviour. Computers in Human Behavior, 75, 547-559. doi:10.1016/j.chb.2017.05.038

Urcuqui. (2018). Malicious and Benign Websites. Retrieved from <https://www.kaggle.com/xwolf12/malicious-and-benign-websites>

Xing, R., Li, J., & Jing, T. (2011). Design and Implementation for Malicious Links Detection System Based on Security Relevance of Webpage Script Text. International Journal of Computer Network and Information Security, 3(3), 41-47. doi:10.5815/ijcnis.2011.03.06

Yearwood, J., Mammadov, M., & Webb, D. (2011). Profiling phishing activity based on hyperlinks extracted from phishing emails. Social Network Analysis and Mining, 2(1), 5-16. doi:10.1007/s13278-011-0031-y

**APPENDIX for Auction Data**

|  |  |  |
| --- | --- | --- |
| **Feature Variable** | **Data Type** | **Description** |
| Title | String | The name of artworks |
| Hammer Price | Int | The final price to be sold at the auction |
| High estimate price | Int | The highest price estimated to sold at the auction |
| Medium | Int | The materials the painting drew on |
| Low estimate price | Int | The lowest price estimated to sold at the auction |
| Signature | String | The position of signature on the paining |
| Create date | Int | The number of bytes transferred |
| Auction Lot | Int | The number of bids on the painting |
| Auction house | String | The institute held the auction |
| Auction Date | Int | The date held the auction of the painting |
| Width | Float | The width of the painting |
| Height | Float | The height of the painting |
| Sold Status | String | The status of being sold or not |