Milestone 3 Report

Current State of the project

We have gathered data for paintings with artists' names and dimensions and begun training models with these data. Overall, we have collected over 24,000 images of paintings, about two thirds of which we have also acquired the artists' names and dimensions. To train with data that includes names and dimensions on top of image data, we have modified our models' input to include an encoding of the artist's name and the painting's dimensions. Our models improved on accuracy. Among the paintings data that we have collected, a little less than half have a very similar price, so our models learned to predict one particular price for every painting. We solved this problem by down sampling data from that price group.

For the website component, we have added a drawable canvas for users to make their own artwork. We also added a dimension input box for users to type in.

Feature Changes

We do not plan to have feature changes.

Current Challenges

The models that train on data with additional information regarding size and artist, still suffer from significant loss despite outperforming models trained on data without this information. One option for addressing this problem would be to break the continuous price range into categories much like we have done in the past with our simpler dataset that contained only images and prices. This showed promise in the past and could likely help in this case as well. Additionally we could clean the dataset to contain only artists that we have collected more than x number of art pieces from. Currently, our dataset contains around 5000+ unique artists, some of which have very few art pieces making it hard for the model to learn anything about that specific artist which will hurt its overall loss. Additionally, having a list of 5000+ artists on our website to choose from may overwhelm the user and may seem pointless.

We worked on the idea of breaking the continuous price range into categories more with our expanded dataset. However, the accuracy we obtained was not as good as our other regression models. It could be because our models were not accurate enough in the first place, and adding another source of error only further decreased their performance.

Another problem with training with additional information is that when users upload images or use their own drawings, it is difficult to generate an input for our model. Since our model input requires an encoding where only a certain number of artists can be encoded, using any other artist's artwork may hurt our model's performance. To overcome this we could allow the user to impersonate a given artist or we could use our non-expanded model to evaluate it.

Tasks Done and Underway by Each Member

Dylan:

- Scraped some additional data for the expanded dataset
- Created model that allows for expanded dataset input
- Trained model on different parameters

Eva:

- Gather additional data with artists and dimensions
- Improve and add more features of the web app
- Successfully host the current app on Heroku (https://artperdict.herokuapp.com/)
- Debug the required files for hosting on Heroku

Jane:

- Scraped additional data with artists and dimensions
- Trained model on subsets of the dataset and on different parameters

Jared:

- Organized and analyzed data
- Trained models using newly gathered data
- Added the drawable canvas on the website