Building Models From Paint
DS4002 Case Study

Background:

Autumn Rhythm by Jackson Pollock

*Walking in the Mis*t by Pol Ledent

As Leonardo DaVinci said, "Painting is poetry that is seen rather than felt". The price of a painting is not always seen. Pricing for paintings is a nebulous subject, debated even among experts. Some paintings are worth exorbitant amounts of money. *Autumn Rhythm*, which showcases a novel drip-paint technique, is estimated at $20,000 dollars. On the other hand, *Walking in the Mist* is worth $1,360 - a small fraction of first price. Therefore, estimating the price of a painting is not as straight-forward as it may seem. Knowing this, the National Gallery of Art wants to come up with a model to predict the general price of a painting given various factors like artist, complexity, location, etc. Beyond the basic information of paintings, the National Gallery of Art is also looking into various image analysis techniques to find any trends with the images themselves and price. This is where you come in.

Task: Your task for this project is to build a prediction model that estimates the prices of paintings, given various predictors/variables. There is a lot of leeway in terms of which model to use (multiple linear regression, decision tree regression, random forest, etc.), so the choice is up to you. After building your model, you will make a GitHub Repository to store the dataset you used for the model and the code you used.

What you will have: You will be provided with a code file and 2 cleaned data sets. The code provided utilizes multiple linear regression in R. Both files will be located in the GitHub (along with these documents).

General notes: You do not have to use the same variables as the code. In fact, part of the project is to independently analyze specific variables that are important for predicting painting prices. Also, you are free to add onto the dataset with other information you may find through image analysis. There will be no accuracy goal to reach since real-world data can be messy and uninformative. However, do think why your model is behaving the way it is. Other than that, good luck and have fun!

References:

- General information about art prices:

S. Koegler, "How is Fine Art Priced in Galleries?" *Medium.com,* Jan. 29, 2022. [Online]. Available: https://medium.com/@scottkoegler/how-is-fine-art-priced-in-galleries-d0bea2a99da0#:~:text=Art%20prices%20are%20notoriously%20opaque%3B%20even%20experts%20can't,by%20other%20people%20who%20like%20what%20you%20make. [Accessed Dec. 6, 2023].

- General information about MLR:

Zach, "How to Plot Multiple Linear Regression Results in R," *Statology.com,* Dec. 23, 2020. [Online]. Available: https://www.statology.org/plot-multiple-linear-regression-in-r/. [Accessed Dec. 6, 2023]

- General information about decision tree regression:

AnkanDas22, "Python | Decision Tree Regression using sklearn," Jan. 11, 2023. [Online]. Available: https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/. [Accessed Dec. 6, 2023]

Image Citations:

Pollock, Jackson. "Autumn Rhythm." *Artsper,* 30 Sep. 2019, https://blog.artsper.com/en/get-inspired/25-works-of-contemporary-art-you-need-to-see/

Ledent, Pol. "Walking in the mist." *Saatchi*, 2023, https://www.saatchiart.com/art/Painting-Walking-in-the-mist/9021/10277541/view

Building Models From Paint Rubric

General Description:

You will build a predictive model to estimate painting, using the dataset given and any other additional information you may find.

Why am I doing this?

This assignment will give a basic introduction to common data analytic methods like multiple linear regression and/or other machine learning algorithms. Additionally, you will become familiar with GitHub, which is a useful tool in data science, CS, and other tech fields. Finally, you will be more familiar with model evaluation metrics when looking at your final model.

What am I going to do?

The first step is to read and understand what your task is, along with the rubric. Next, go to the GitHub repository (link in the reference page) to understand the dataset and the variables in it. Once you have understood the variables and dataset, begin brainstorming which variables are important for predicting price. You can always add to the dataset if you want to include other variables. After that, begin prepping your data and building your model. After building your model, make sure to look at the evaluation metrics to determine how your model is doing, and think of reasons why your model is behaving as such.

Tips for success:

- *Ask for help when you need it:*
  - Your professor/TA are here to help you. So, don't be afraid to ask for help.
- *Research online:*
  - Some basic code and resources are linked in the GitHub, but you're probably not going to find everything you need code-wise, so researching online is a great way to find what you need/didn't know you need.
- *Start general, then go granular:*
  - Instead of diving straight into the code, have a general idea of what you want to do (which original variables to include, what new information to add, etc.).
- *One step at a time:*
  - It helps to break things down into sections and go through them one at a time.

How will I know I have succeeded:

Spec Category	Spec Details
Formatting	● GitHub Repository
GitHub Repository	● Include:

	- README.md files (describe the project context, model-building process, and the dataset you used) - Data folder (the dataset you used to build the model) - SRC folder (the code you used to build the model) - Figures (Exploratory data analysis or post-model evaluation visualizations, if needed) - License (terms if others want to use your work, Use MIT format)