**Adam Henze**

**Sambriddhi Mainali**

**CompSci4200**

# Final Project Written Report

## Problem Definition

The goal of this project was to determine the correlation between weighted indicators of stock portfolio options using a multiple regression learning model. The correlations of these weighted indicators will be used to determine to potential performance of a stock over time.

## Background and Motivation for the problem

Although statistics may not be my favorite form of arithmetic, I was motivated to study machine learning in order to further investigate the applications towards financial markets and the interpretations of large data sets for easy visualization and understanding. I was motivated to undertake a financial based project in order to begin to develop code suited for use with financial markets and systems.

## Data Source

I will be using data available in the UCI data repository from data set https://archive.ics.uci.edu/dataset/390/stock+portfolio+performance. There are six predictor variables, book value to price ratio, sales to price ratio, return on equity, return rate in the last quarter, market capitalization, and systematic risk. There are six potential target values for investors to determine their stock options by; Annualized return rate, Excess return rate, Systematic risk, Total risk, Absolute winning rate, Relative winning rate.

## Machine Learning Model

### Algorithm Description

In order to understand the effect of each weighted performance indicator on the dependent variables of each stock we will be performing an unsupervised machine learning model to allow the algorithms to randomly parse the data to find correlations between dependent and independent variables. First we need to clean the data and correlate it with independent values. We parse the relevant columns of independent values and perform a correlation analysis with the relevant dependent variable to find the coefficients of the weighted performance indicators on each dependent variable.

### Model Training

Our machine learning algorithm was trained by using a linear regression fit model on an 80/20 split between training and testing data. Using randomly selected data members and an

unsupervised machine learning model we are able to parse the data and perform predictive analysis using incomplete data members to test the validity of our model.

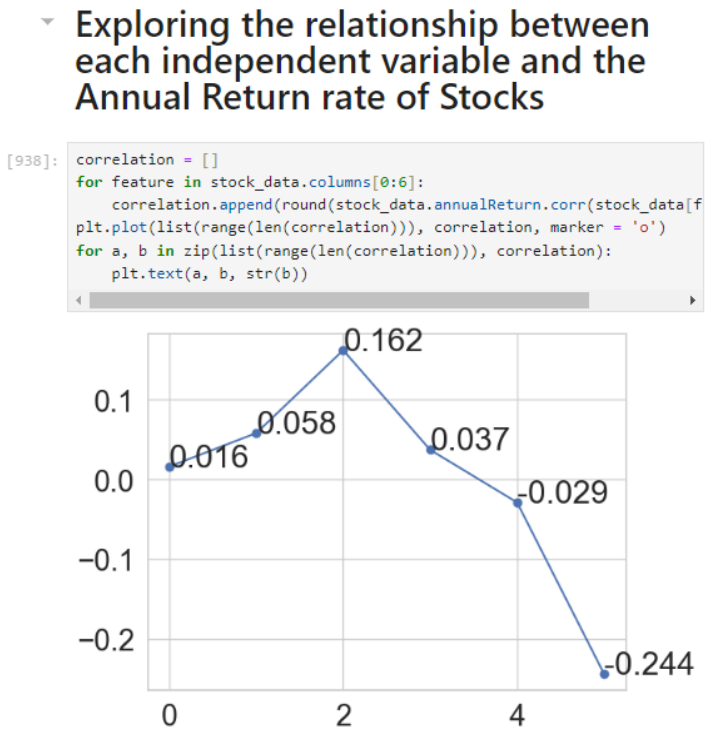
**Evaluation**

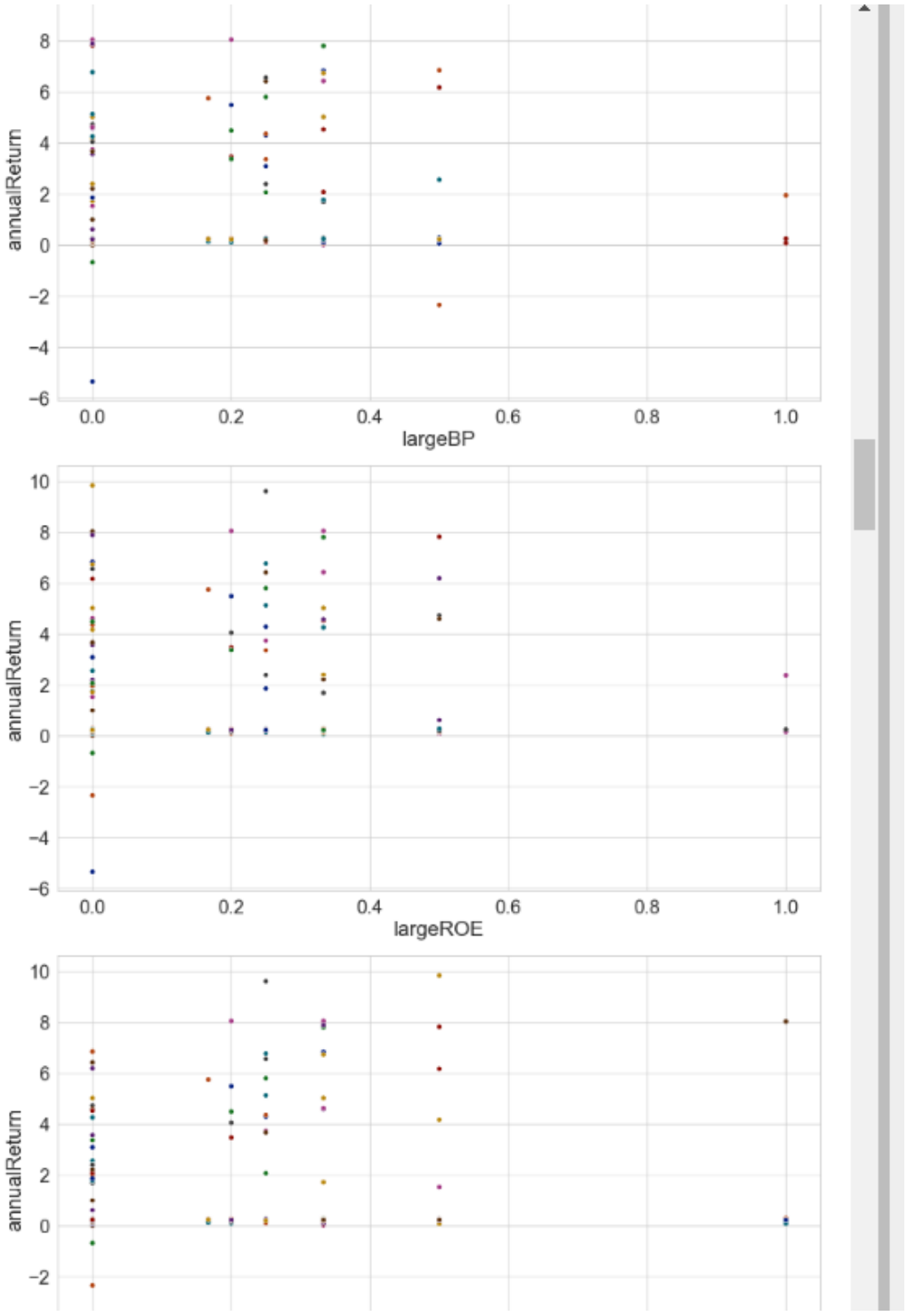
The machine learning algorithm was able to successfully divide and correlate the independent variables of the data set to coefficient values of their effects on the dependent variable of Annual Return.

**Results**

Our program was able to successfully gather data and perform visual analysis on the data set. The visualizations allowed us to see clear correlations and non-correlations between independent variables and dependent variables. We were able to use these correlations to examine a large data set and begin to train a machine learning model with a randomly selected population of test subjects and training subjects.

The linear regression model used by the unsupervised machine learning algorithm was unfortunately not able to make meaningful predictions towards individual stock performances based on the weighted indicators alone. This could be due to a small data set, requiring more data to conclusively begin to assess performance indicators. This could also be due to the limited information available form the performance indicators. We can see in the visualizations the certain performance indicators do not lead to a strong correlation between stock performance and the indicator value.






**Conclusion**

In conclusion, this project aimed to explore the correlation between weighted indicators of stock portfolio options through the application of a multiple regression machine learning

model. Motivated by a desire to delve into the intersection of machine learning and financial markets, the study utilized data from the UCI data repository, encompassing predictor variables such as book value to price ratio, sales to price ratio, return on equity, return rate in the last quarter, market capitalization, and systematic risk, with six potential target values for investors.

The machine learning model employed an unsupervised approach to randomly parse the data, seeking correlations between independent and dependent variables. The algorithm underwent training via linear regression on an 80/20 split between training and testing data. Despite successful data gathering and visual analysis, the linear regression model did not yield meaningful predictions for individual stock performances based on the weighted indicators alone.

Several factors may contribute to this limitation, including the potential need for a larger dataset or the inherent constraints of the available performance indicators. Visualizations highlighted instances where certain performance indicators did not exhibit a strong correlation with stock performance. This underscores the complexity of the relationship between financial indicators and stock outcomes, suggesting that a more nuanced and comprehensive approach may be necessary for accurate predictions.

In future endeavors, expanding the dataset and incorporating additional relevant factors could enhance the model's predictive capabilities. Despite the current limitations, this project serves as a valuable foundation for further exploration at the intersection of machine learning and financial analysis, offering insights into the challenges and opportunities within this dynamic field.