**Stock Movement Prediction Using Market Features and Unsupervised Learning**

**Carlos Semeho Edorh (MSCS, 2025)**

**Zachary Coates**

**(MSCS, 2025)**

**Benjamin Northrop**

**(MSCS, 2025)**

**Professor Mohammad Toutiaee  
CS6220: Data Mining — Summer 2025**

A pie chart with different colored circles

AI-generated content may be incorrect.**Introduction**

The Standard & Poor’s 500 Index (S&P 500) is a list of the 500 largest publicly traded companies in the United States, considered to broadly represent the country’s overall economy. Figure 1 shows the various market sectors that comprise the S&P 500. Virtually every mutual fund compares its performance against the S&P 500 as a benchmark, and the goal of every financial advisor and brokerage firm is to *beat* the return of the S&P 500.

We attempt to use various techniques to evaluate the historical performance each ticker within the S&P 500, analyze it, and produce a prediction of little movement, or strong movement either up or down. By adjusting the features we analyze and our methodology, we are able to achieve 70% accuracy in our test sets, indicating our model is able to glean valuable momentum information regarding the present day market.

Figure 1.

The benefit of finding such information is that traders will be able to leverage the model to make better investment decisions and would build wealth more quickly than the “buy and hold” technique often used in index fund focused strategies. Our model assumes that markets have both high frequency fluctuations and lower frequency trends, and that both may be captured by evaluating the daily returns over various time periods.

Our target variable is the 1 day return of a ticker, based off of features that include: previous return data, volatility, RSI, and MACD.

**Methods**

**A group of different colored squares

AI-generated content may be incorrect.**

Figure 2. The program pipeline

The program is built as a pipeline that provides outputs at each step, allowing the user to choose the model that performs the best. The first step is checking for a current feature set, or downloading a new one. If building a new data set, the program uses the YFinance library to interface with the Yahoo Finance API to the data for all tickers within the S&P 500. The list of tickers comes from Wikipedia. The program then organizes each data set into a Pandas dataframe before writing it all to a csv. This csv is saved in the data folder.

The first step in data analysis is via PCA analysis. We load the csv and perform the final data cleaning techniques. Our model uses a maximum time horizon of 120 days to capture the overall trend of the model in the past 4 months. If a ticker has not been active for the that long, that data is blank. To ensure missing data does not skew the model, we drop all rows with missing data. This ensures all data is for tickers that have been active for at least 4 months. We also separate the file into a features set and the target set. The target is the 1 day return percentage, and the features are all of the others. We drop columns that cannot contribute, such as the date of the data, the ticker label, the actual adjusted close, etc. After scaling the data, we perform a fit and transform on the dataset and save it to a .csv file. Rather than determine the exact number of components, we leverage sklearn’s ability to capture 95% of the variance.

The baseline model uses a logistic regression to determine

**Dataset**

Our data set is the historical financial data of each of the 503 stocks that currently comprise the S&P 500. The python library Y Finance interfaces with Yahoo! Finance’s API and downloads a wealth of stock information requested by the user. In our model, we divide each stock into several feature sets: simple returns, MACDs, and RSI. We use the daily return percentage of various horizons: 1-5, 20, 60, and 120 day horizons. For training and testing historical data, the 1-day horizon becomes the target value. Mean Average Cumulative Distribution (MACD) computers a weighted average between a short term time frame and a long term time frame, giving a sense of short term movement in relation to long term trends. The time frames we use for this are 12 days for short, 26 days for long term, and 9 day signal. Volatility is an indication of how much the stock may fluctuate in a given period. Higher volatility indicates large price swings, thus higher risk and reward. The RSI is a metric to indicate whether or not a stock is overvalued or undervalued, and often used to indicate a buy or sell signal of a feature. We use RSI values of the previous 14 days.

**Results**

**Discussion**

**References**

**Appendix**

**Statement of Contributions**

1. **Introduction:**Provide a short background of the project (e.g., what kind of question is to be answered, and why this is of interest). Provide a short non-technical summary of your analysis.
2. **Methods:**Summarize the methodology used in your analysis and use mathematical formulas and notations.
3. **Dataset:** Identify a dataset to study, and describe literature related to the problem you are studying. If you are using an existing dataset, where did it come from and how was it used? What other similar datasets have been studied in the past and how?
4. **Results:**Show the results of your analysis.  Summarize the results with a small number of the most important figures or tables and keep the description short.
5. **Discussion:**What are the state-of-the-art (SOTA) methods currently employed to study this type of analysis? Are the conclusions from existing work similar to or different from your own findings? Why did the proposed model succeed or why did others fail (or if it failed, why did it fail)? What did you learn from this analysis? What additional steps could be potentially performed to improve your analysis?
6. **References:**Please only add references that are explicitly used in the text. Make sure that you cite the sources of data, and the associated claims. Use a consistent format and numbering scheme. Example:

      [1] Bob Smith, John Doe. My amazing method. In \_Proceedings of WWW 2024\_, Lyon France, 2024.

1. **Appendix:**Add links to the code repository, plots, and other relevant technical details that will help evaluate your work.
2. **Statement of contributions:**If you are working in a group of 2 or 3, your report should contain a short section at the end, stating that all group members contributed equally, and describing the contributions of each group member. You do not need to include this section if you are working alone. This section does not count towards the page limit.

**Note:**

* Your report should also include a link to a public GitHub repository containing the code used to produce your results. If you would like to not make your code public for any reason, you must instead submit your code as a zip file on Canvas.
* Your report should be written as a formal academic article. There should be no slang or jargon. Any claims you make should either be common knowledge (i.e. something we learned in the course), supported by a citation, or supported by your data/results. There should be no spelling or grammar mistakes.
* One submission per group is enough.