Course Project Checkpoint

CSCI-4502

William Ruiz
Computer Science
University of Colorado Boulder
Boulder Colorado USA
wiru2611@colorado.edu

Kasper Seglem
Computer Science
University of Colorado Boulder
Boulder Colorado USA
kase1828@colorado.edu

Angelo Vacca
Computer Science
University of Colorado Boulder
Boulder Colorado USA
anva5757@colorado.edu

INTRODUCTION

For our project, we are analyzing the stock history of the current S&P 500 companies. We hope to find patterns and determine correlation between stocks, using this information to make predictions. As an example, stocks in the same industry may often go in the same direction. This data would be similar to itemsets, where the stocks with similar movement would be in the same set. Already, we have gathered much data about correlations between these companies. Therefore, our focus going forward is on using this data in predictive models and evaluating the results.

KEYWORDS

Finance, stocks, correlation, S&P 500, predictions, trading, algorithms.

RELATED WORK

1 Hudson River Trading [1]

Hudson River Trading is a firm located in New York that specializes in algorithmic trading of stocks through rigorous analyses and attempting to predict future price movements based on live data. This firm utilizes their trading algorithms to invest your money in order to beat the S&P 500 average.

2 J.A.R.V.I.S [2]

JARVIS is similar to Hudson River Trading, however the software utilized runs on top of the TradingView platform to give the user a more hands-on experience versus a company doing the trading for you.

We will try to take inspiration from both of these algorithms, while also developing our own methods. However, we hope to mainly rely on our own data findings.

3 Open Source Projects

There are many open source projects in this field and we would like to study these in order to learn from them and compare them against our own. This gives us a chance to test our predictive model against real world examples from other people. Compared to the other examples, open source projects will be easier to access and will also be more akin to our code.

PROPOSED WORK

1 The Dataset

We will be using a dataset containing the historical data for each of the current S&P 500 companies. It includes six different values, all of which may be used for our algorithm. The dataset is publicly available at:

https://www.kaggle.com/datasets/camnugent/sandp500?resource=download.[3]

Kaggle is a subsidiary of Google, where people can post datasets. It was made primarily for data scientists and machine learning practitioners.

It contains data from the past five years. It is in comma separated value format and contains the following columns for each day of the market:

- Date - in format: yy-mm-dd
- Open - price of the stock at market open (this is NYSE data so all in USD)
- High - Highest price reached in the day
- Low Close - Lowest price reached in the day
- Volume - Number of shares traded
- Name - the stock's ticker name

We will likely start with the most recent activity and work our way backwards.

2 Correlation between stocks

One idea we have is to study the correlation between individual stocks. If a strong correlation is proven, we can expect the stocks to perform similarly and use that information to make predictions. This is something that can be seen with stocks in the same industry. It is often the case that these stocks move together, or at least have some shared qualities.

3. Predictive Algorithm

The main goal for the project is to create a predictive model for the stock market. This could include using a stocks own historical data or using the data of its closely correlated stocks. We may also include elements from other algorithms if we decide they may be useful.

PROGRESS UPDATE

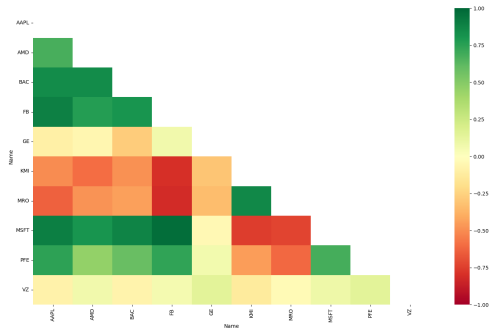
We have been able to develop functions that allow the processing of raw CSV data into a refined dataframe structure utilizing pandas and numpy. It is important to process this data as in order to utilize it for correlation analysis it needs to be formatted in a different way in order to perform correlation calculations on the stocks. We are using the Pearson method of correlation and the function allows you to input a metric (high, low, open, close, volume) and it will calculate the correlations between the stocks and format the dataframe accordingly. From there the dataframe can be inputted into another function with a few parameters like starting index and number of stocks to generate a correlation heatmap like you can see below. We found seaborn to be the most powerful tool in developing these functions as it has easy to use plot generation and color schemes. From here we plan to develop functions that allow easy generation of top/bottom 25 companies of

a certain metric and line plots that allow us to see change in sectors over time. Along with that we have found more data that helps describe the S&P-500 data we already have with information like their sector and the full company name. [4]

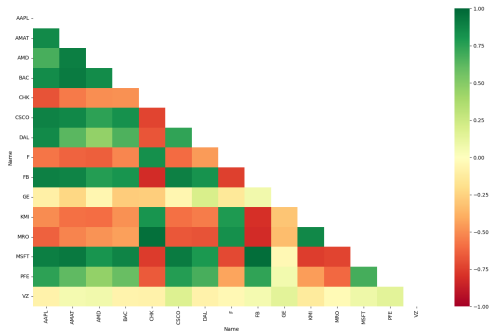
INITIAL ANALYSIS

Using the python tool seaborn, we were able to get correlations between stocks and display it in heatmaps. We can also store this data for further use in our predictive models going forward. The following are stock correlations between the current S&P 500 companies:

Top 10

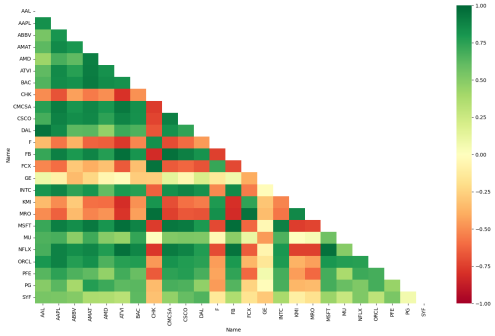


Top 15

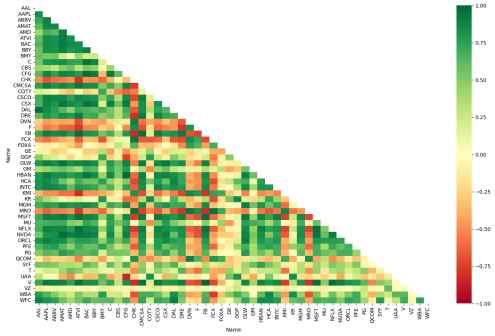


As expected, we found that companies in the same industries are more likely to be correlated. For example, Microsoft has strong correlation to Apple and weak correlation to Ford.

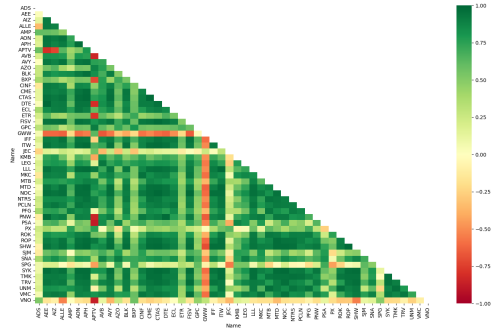
Top 25



Top 50



Bottom 50



We also decided to check the bottom 50 companies in the S&P 500 to see if there was any difference in the heatmap. It was clear that these are much more correlated to each other than the top 50. We hope to explore this going forward.

REMAINING WORK

Now that we have a dataset and working demos, we plan to mine further into the dataset, identifying correlations along with potentially clustering similar stocks. An example of some of the data we plan to correlate is data within a singular year of the dataset. With these correlations, we plan to start building a predictive model that will predict market changes along with specific stock changes over time. While this model will start with our 2013-2018 dataset, we hope the insights we find will grant us understanding as to which factors affect the market the most and extend to other datasets we plan to use in the future. We also hope these insights can find some success when it comes to trading on the market, although we do understand that this most likely will not be the case. Additional work we would like to do would be to host the application on Heroku or a similar platform.

EVALUATION

We can evaluate our work by analyzing our predictions and whether they correlate/match stock changes or patterns for that day. We can also compare our model with other models such as Hudson River Trading or J.A.R.V.I.S and see how our model compares. While we do not expect to have accuracy close to matching that of these professional made models, they can give us a good baseline as to what a well made model is able to do and predict. Another way we can evaluate our work is by comparing our predictions to randomness. This can give us insight into knowing whether our model can actually predict trends and changes in the stock market, and is not simply a random algorithm.

Our metrics for this evaluation will be the accuracy of our predictions. More specifically, it will be the percentage change predicted by the algorithm, compared to the percentage change of the actual stock being tested. Percent error ranges will also be used when evaluating our model, and statistical tools like standard deviation can tell us how off our predictions are. Comparing our error ranges to that of randomness and other models will also allow us to evaluate how much better or worse our predictions are compared to other approaches and algorithms.

MILESTONES

1 **Week 7 - COMPLETE**

Collect past and current stock data from the S&P 500. This includes finding an up to date dataset with the history of each stock in the S&P.

2 **Week 10 - COMPLETE**

Study developed trading algorithms and models used to help predict stock movements. This includes learning more about the Hudson River Trading algorithm and the J.A.R.V.I.S. algorithm.

3 **Week 12 - COMPLETE**

Project checkpoint report. Put together a report to document current progress.

4 **Week 12 - COMPLETE**

Find correlations between the top S&P 500 stocks. Find correlations between the bottom S&P 500 stocks.

5 **Week 13**

Study and mine data in order to gain a better understanding of historical stock data by finding trends, patterns and outliers in movements. Use the S&P 500 data and begin to mine patterns and correlations.

6 **Week 13**

Cluster similar/correlated data and stocks.

7 **Week 15**

Use data from the previous milestone to create a rudimentary predictive model/algorithm based on patterns and trends from mined data that allows fictitious stock trading to happen.

8 **Week 16**

Testing, final project and final report. Study the accuracy of our new algorithm and document it in the final report.

REFERENCES

[1] Hudson River Trading - HRT. n.d. *Home » Hudson River Trading - HRT*. [online] Available at: <https://www.hudsonrivertrading.com/> [Accessed 6 October 2022].

[2] Jarvis-algo.com. n.d. *J.A.R.V.I.S - trading algorithm*. [online] Available at: <http://www.jarvis-algo.com/> [Accessed 6 October 2022].

[3] Nugent, C., 2018. *S&P 500 stock data*. [online] Kaggle.com. Available at: <https://www.kaggle.com/datasets/camnugent/sandp500?resource=download> [Accessed 6 October 2022].

[4]Hub, D. (2017) *S&P 500 Companies with Financial Information*, *Datahub*. Available at: https://datahub.io/core/s-and-p-500-companies#python [Accessed: November 15, 2022].