CSCI 4502 Group 4: Stock Market Analysis - Part 3

Stephen Kay[*]
Rachel Mamich[†]
Levi Nickerson[‡]
Brandon Rajkowski[§]
Stephen.Kay@colorado.edu
rachel.mamich@colorado.edu
Levi.Nickerson@colorado.edu
Brandon.Rajkowski@colorado.edu
University of Colorado at Boulder

ACM Reference Format:
Stephen Kay, Rachel Mamich, Levi Nickerson, and Brandon Rajkowski. 2019. CSCI 4502 Group 4: Stock Market Analysis - Part 3
.

1 PROBLEM STATEMENT/MOTIVATION

The stock market has historically been unpredictable. Some people make a fortune while others lose everything. The goal of this project is to analyze historical stock market prices and trading volume in order to develop an investment strategy. Hopefully this investment strategy will serve as a guide to smart investments.

2 LITERATURE SURVEY

Currently Decision trees and artificial neural networks are the two most popular formats for using data mining to manipulate the stock market with 22 and 17 percent of the current market share respectively (Liao et al. 2012). Specifically for decision trees it is important to have a well pruned tree so there must be time that is put into a training set. Pruned trees tend to have smaller error rates than methods such as k-nearest neighbors or even artificial neural networks (Kian and Rasheed 2006). One analysis form is usually not enough to beat the stock market and thus many different ensembles are used for data mining analysis when trying to beat the stock market. As mentioned previously one other popular data mining techniques is that of artificial neural networks. These data sets are very tedious to implement but when done properly can have very powerful results. A study by (Guresen et al 2011.) found that the proper implementation of an artificial neural network can correctly predict whether the final NASDAQ score will be up or down. This particular neural network predicted a final score of 1737.70, which was frighteningly similar to the actual NASDAQ score of 1747.17. However, just like any data mining technique this evaluation technique is incredibly complex and is always being refined (Guresen et al 2011.)

3 PROPOSED WORK

The proposed work for this analysis can be broken down into three major steps: data cleaning, data integration, and data processing. During data cleaning, all of the data sets will be conformed into a consistent format and the actual data points will be checked for missing and inconsistent values. After the data is clean, it can then be integrated into a single source that is ready for processing. Finally, the data will be processed using various data mining techniques to expose an investment strategy.

Before any data mining can be conducted, the data will have to be cleaned into a consistent format. In this analysis, several different data sets containing stock market and ETF data will be used. Each data set has a different format for describing the timestamp of when the data was collected. In order to be able to compare timelines between the data sets, all of the timestamps will have to be converted into a consistent format. All of the timestamps from each data set will be converted into the standard ISO 8601 timestamp format. This can be accomplished by running a script over all of the data sets that parses the existing timestamp and replaces it with the ISO 8601 timestamp. In some instances, data may not have been collected over a time period even though the market was still active. Any missing data points will be calculated and filled in by interpolating the values of the two nearest existing data points. This is a reasonable method for short time frames of missing data since historically stock prices are relatively stable over short time periods. Finally, stock splits will need to be accounted for in order to maintain consistent data. When a stock splits, its price is cut in half and shareholders are then given double the number of shares. Therefore the market capitalization of the stock remains the same, however the price is cut in half and the number of

[*]SID: 109202680
[†]SID: 104786655
[‡]SID: 109340569
[§]SID: 101279173

shares outstanding is doubled. When conducting an analysis on just the pricing and volume data, the reduction in price due to a split would appear to look like a huge loss for the day. However, no actual losses have been incurred. All of the pricing data will be converted into a split-adjusted share price.

Once the data has been cleaned, it can be integrated into a single data source. Not only are there several data sets being used for this analysis, some of the data sets have multiple files within them. All of the separate files across all of the data sets will need to be combined into a single source. This can be accomplished by running a script over all of the files and writing the data to a single source. Furthermore, during this process the timestamp needs to be considered in order to align the data temporally.

After organizing the data temporally and into a single source, the data can be processed with various data mining techniques. The first technique that will be used is a basket of goods analysis using the Apriori algorithm. During this analysis, volume will be used as the support indicator and the minimum support value will likely have to be adjusted to obtain baskets of comprehensible sizes. The goal of this analysis is to determine what stocks appear to be bought and sold together in high frequencies. Frequent pattern mining techniques will be used to spot trends in pricing across stocks. This analysis will answer questions regarding the symmetry and asymmetry of price movements between stocks over a time period. The frequent pattern mining should also expose any price patterns that occur before large price swings; these patterns can serve as buy and sell indicators for the investment strategy. A cluster analysis over all of the individual stocks can be performed over a time period to group stocks into clusters based on their price and volume. After the clusters have been determined, the overall rate of return of the cluster can be calculated. Clusters with higher rates of return can indicate stocks that not only performed better, but also had very similar price and volume data. Finally, a decision tree will be created based on the price and volume that can be used as a "buy", "no buy" decision maker when considering new stocks to add to a portfolio.

4 DATA SET

There are a few different data sets that the group has already gathered. The first is Kaggle US Stocks and ETFs. This data set can be found at:
https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs. This data set contains daily stock market values for price and volume for the NASDAQ, and NYSE markets as well as ETF's for the same markets. There are over 7,000 different stocks accounted for beginning in different years and going until early 2017.

Another useful data set for analyzing the stock market is NASDAQ Historical Quotes which can be found at:
https://www.nasdaq.com/quotes/historical-quotes.aspx. This website allows users to pull information on specific stocks for different spans of time from the last 3 months to the last 10 years.

The last data set that the group plans to use is NYSE Historical Quotes which can be found at:
https://www.kaggle.com/dgawlik/nyse. This data set contains data for the prices in the NYSE market. There are separate files for prices accounting for splits and for prices without including splits. This data spans 2010 to 2016. This data set also contains metrics extracted from annual SEC 10K fillings (2012-2016).

5 EVALUATION METHODS

The goal of this analysis is to develop an investment strategy. In order to judge the validity of the strategy its performance will be compared to that of the S&P 500 index. A mock portfolio will be created based on the investment strategy. The portfolio's returns will be tracked for several weeks and compared to the S&P 500 returns. At the end of the evaluation period, the final return of the portfolio will be compared to the final return of the S&P 500. Outperforming the S&P 500 indicates an above average portfolio and a successful investment strategy while falling short of the S&P 500 indicates a poor portfolio and strategy. By tracking the daily performance of the portfolio a winning percentage can be calculated which indicates the percentage of days that the portfolio beat the S&P 500. This metric along with the final return of the portfolio can give investors an idea of the short term and long term performance of the underlying investment strategy.

6 TOOLS

The primary tools that will be used throughout the project are Python, Orange, and Druid IO. Python will be used to modify and maintain the data. The Pandas library will be used to build the data into a database. Python and specifically the Pandas library were chosen because they are capable of holding a database with millions of rows. The Numpy and Scipy libraries will then be used to perform modifications and computations regarding the data. These libraries will be important for the preprocessing of the data. They will be used to clean the data and fill in values for missing data. The Matplotlib library will be used for data visualization. Orange is an open source data mining tool. The primary functionality of the tool is the ability to join two data sets. We will be utilizing multiple data sets and it is important that we can combine the sets if needed. Orange also has visualization features that could be used to show results from the data mining. Druid IO is another open source tool we will

be using. The primary use for Druid IO will be performing OLAP queries on the database. Druid IO is capable of taking on large data sets and performing queries with limited delay. All of these tools together should provide the functionality we require for the project.

7 MILESTONES

The initial milestones for the project are as follows. We intend to have the data preprocessing done by March 24 for the Progress Report. This is a crucial part of the project and we may dedicate more time to it if we encounter unforeseen problems. We also intend to have begun data analysis by that date. The sooner we can begin the analysis the sooner we can develop our model and start collecting results. For the Progress Report we will complete the write up by March 23 so that we have all of March 24 to make changes and corrections. Once the data is processed and the initial analysis is complete we can begin to develop our investment strategy. The goal is to complete our strategy by the first or second week in April. This will allow nearly a month of time to collect data regarding our strategy compared to the S&P 500 index. The first week in May will be spent working on the Final Report and the Project Presentation. We intend to have both of the assignments done by May 4. This will give us time on May 5 to have a final look at the assignments and make corrections as needed.

Table 1: Milestones for Group 4, Stock Market Analysis

Date	Milestone
March 11, 2019	Project Proposal (Part 2)
March 23, 2019	Data Preprocessing
March 23, 2019	Progress Report (Part 3)
April 12, 2019	Complete Investment Strategy (Rough)
May 1, 2019	Finalize Investment Strategy
May 5, 2019	Project Final Report (Part 4)
May 5, 2019	Project Presentation (Part 5)

8 MILESTONES COMPLETED

At this point in time, the first three milestones have been completed. This includes the Project Proposal, the Data Preprocessing, and the Progress Report which is the subject of this paper. Although there will likely be continued data preprocessing, the data has been processed to be compatible with Orange. The first major step of preprocessing was to convert the data files into a format compatible with Orange. The data set provided separate files for each stock and ETF in text file form. However, Orange is not compatible with text files. In fact, you cannot load text files into the program at all. In order to prepare the files for Orange, a script was created that converted all of the text files into csv files.

Once the data was compatible with the Orange platform, the individual files were combined. Another script was created that would allow for the combining of arbitrary files into one file. The Python script takes in two csv files and returns one csv file with all of the combined data sorted by date. Python, specifically the Pandas library, was selected to join the data because it was the most efficient method to correctly manipulate the data. The only indicator of stock association with the data is by the file names supplied. In order to combine the data, another attribute column for the tag name needed to be added so that there was no loss of information during the combining. The script pulls the tag name from the file name and appends that information in the newly created column. This allows for much easier data manipulation in Orange since only one file is needed to compare various stocks side-by-side. This allows us to group stocks of interest and utilize Orange to find potential correlations between the stocks. Figure 1 displays a sample image generated from Orange. The image shows the closing costs of Agilent and Alcoa over the lifetime of the stocks. The image shows the unpredictable behavior of stocks that we are attempting to capture and model. The data used to create this image was processed using the scripts described prior. The data was transferred from text format into a csv file and then combined with another file. This new file was then imported into Orange and the scatter plot of Figure 1 was generated.

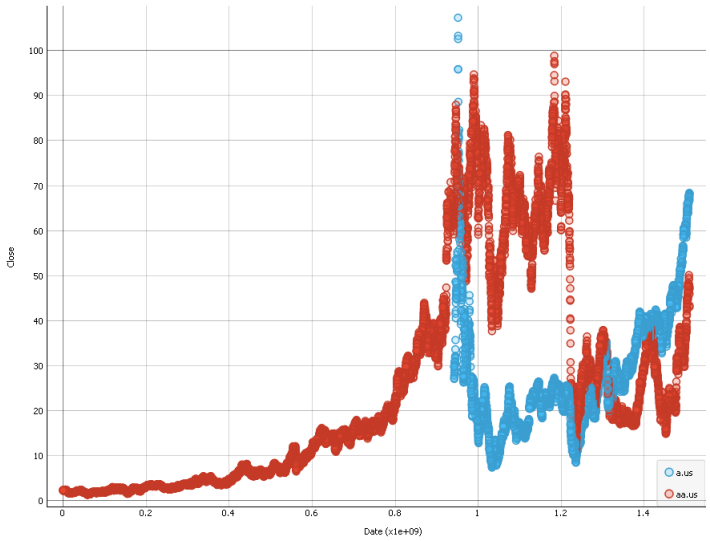

Figure 1: Closing Cost of Stocks Agilent and Alcoa over Time

A key missing attribute that the data set does not provide is the rate of return. It makes sense that the data set does not provide this metric because rate of return is a relative metric. One might want to know rate of return for the day,

month, year, or n-years. To gain access to this metric another script was developed. The python script is able to take a stock, a start date, and an end date and calculate the rate of return for that time period. The rate of return is calculated as the difference between the closing values for the start and end date divided by the closing value of the start date. This will prove to be an invaluable metric when judging the performance of individual and groupings of stocks.

As of right now the data has been processed to an acceptable point. We may wish to further process the data and if we do it will most likely be reducing the data. The data we are using holds all of the historical data for each of the stocks. This provides us with a wealth of information, but it may be detrimental to identifying patterns and trends. We intend to keep all of the data we currently have, but we may be forced to begin reducing the data in an attempt to identify more concrete trends. Other forms of data processing will be used to make the data compatible with any other tools we choose to use. We have not begun using Druid IO for OLAP queries and we might have to perform further processing in order to provide compatible data.

9 MILESTONES

The next milestone is to develop a rough investment strategy based on the data mining results. Therefore this milestone requires some initial knowledge discovery before it can be completed. The Orange platform will be utilized to execute the various data mining techniques outlined in section 3. One of the knowledge discovery goals is to determine which stocks are commonly bought and sold together. Volume is a metric used to tell investors how many shares of a stock have been traded on a given day. By using volume as a metric, the Apriori algorithm can be applied to discover which stocks have similar volumes over time and are thus commonly bought and sold together.

Another goal is to try to find stocks that behave similarly in regards to their price trends. One of the ways to discover this behavior is to perform a correlation analysis on different stocks' prices throughout time. A strong positive correlation would indicate that the stocks' prices tend to behave similarly. A strong negative correlation would indicate that the stocks' prices tend to behave opposite one another. This type of correlation would present a possible investment strategy where one stock could be bought to the hedge the risk of another stock. In another case, investors could use the correlation information in order to place options on stock. Knowing that two stocks are negatively correlated would allow an investor to place short bets on a stock when the negatively correlated stock is seemingly bullish. In addition to these goals that determine how various stocks behave together, there is also potential for knowledge discovery by looking at how an individual stock behaves. Techniques such as decision trees, Bayesian classification, and neural networks could potentially take in individual stock performance data from a prior time period in order to determine a buy or no-buy signal for the present time. Training data can be built from the historical stock data where the input might be the last month of stock performance data, and the output might be a buy/no-buy decision based on the next month's performance. By conducting this training over all of the stocks in the US market, a network can be trained to make the buying decision for an investor.

After the rough investment strategy is developed, it will be tested against the performance of the S&P 500. The performance of the rough strategy is important, but the main goal of the rough strategy to make sure there are no bugs or misunderstanding of the data mining results. If there are any issues with the results, more data mining will be conducted and applied to create a final strategy. This final strategy will be judged on performance solely. The final strategy is successful if it beats the returns of the S&P 500, otherwise it will be considered unsuccessful. Although the strategy might be deemed unsuccessful, the knowledge that was discovered may still be useful to know. Furthermore, due to the natural unpredictability of the stock market, this investment strategy might have performed badly for the first month, but could end up beating the S&P 500 given a longer time period of testing.

All of the data mining results will then be summed up in a final report. This will include any interesting correlations and patterns in the data as well as the results of the buy/no-buy decision machine, and the investment strategy with performance results. All of these items will be presented in a written document similar to this one, but with the extended information.

Finally, a summary of the final report will be turned into a slide show presentation. This presentation will contain all of the high level details and results of the project.

10 RESULTS SO FAR

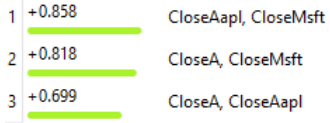

Figure 2: Closing Price Correlations

Before scaling the data mining process up to full scale using the entire data set, a small sub-set of the data set was used to produce proof-of-concept algorithms and plots. One of the potentially interesting metrics is price correlation

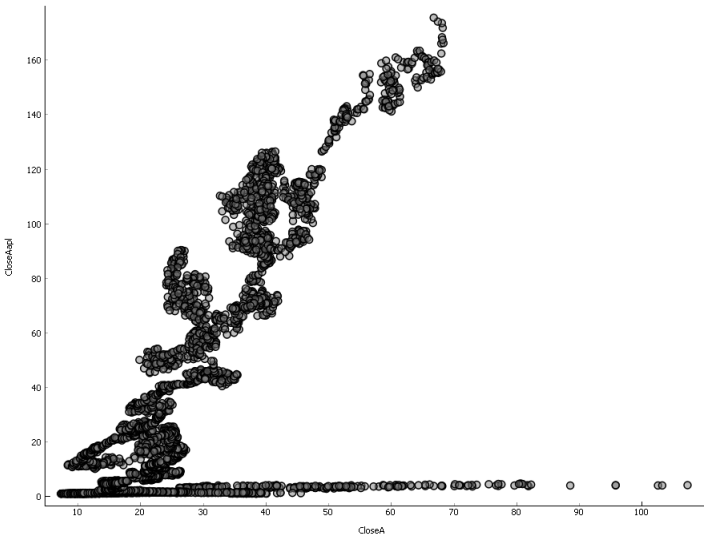

Figure 3: Closing Price of AAPL vs. A

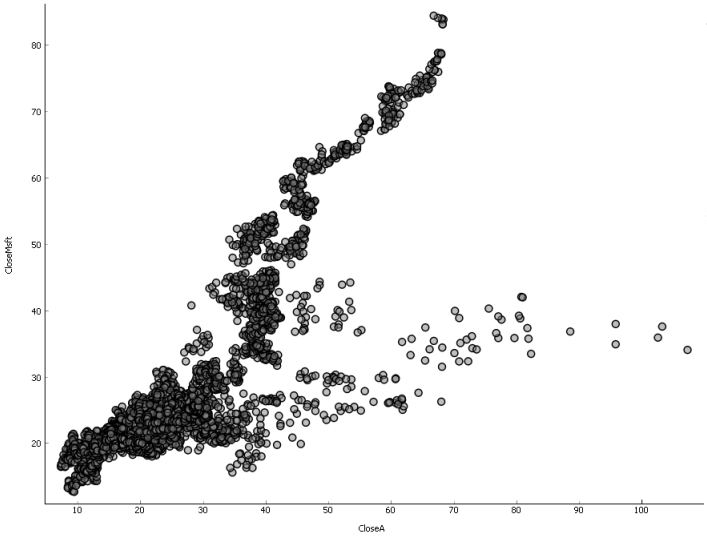

Figure 4: Closing Price of MFST vs. A

between stocks. Knowing which stocks are positively correlated and negatively correlated can help investors make portfolio selection decisions.

As a proof-of-concept, three stocks are analyzed to determine how their prices are correlated between one another. Stock "A" is Agilent, "AAPL" is Apple, and "MSFT" is Microsoft. Before doing any data mining, the pricing data was aligned temporally. The earliest common start date was found between each stock as well as the latest common end date. These two dates formed the time period for the analysis. The closing prices for each stock for each day within this time period was merged into a single file. From there Orange

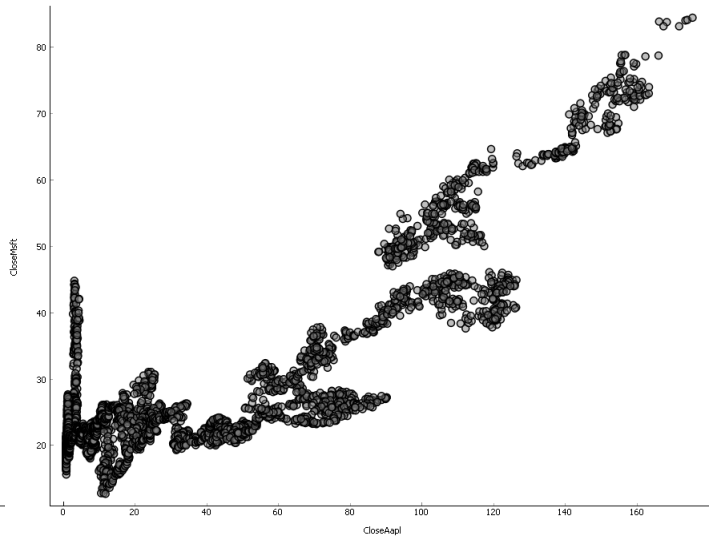

Figure 5: Closing Price of MFST vs. AAPL

was able to calculate the correlation coefficients. Figure 2 shows the calculated correlation coefficients between these three stocks. The highest correlation coefficient is AAPL and MFST. One supporting reason that these stocks are strongly, positively correlated is that they are both considered tech stocks. Usually stocks in the same industry tend to trend very similar.

Sometimes correlation coefficients can be misleading and thus visual confirmation is often required to say two things are correlated. Figures 3, 4, and 5 are scatter plots of each combination of the stocks that were analyzed. Figure 3 shows a strong positive correlation between AAPL and A, however there are many data points that do not follow the correlation and appear to form a straight line. This behavior is explained by noting that stock A had a large takeoff before that of AAPL. During this time AAPL was stagnant while A underwent a boom.

Figure 4 reveals that A and MSFT might not be as correlated as the correlation coefficient suggests. The data in this figure appear to have some correlation, but have a much greater spread than that of Figure 3 and Figure 5. Finally, Figure 5 reveals a strong postive correlation even though there is an apparent vertical line in the data. As in Figure 3, this can be explained by knowing MSFT had a boom prior to AAPL having a boom.

Since it is possible to calculate correlation coefficients at small scale, this process can be extended to calculate correlation coefficients between every stock in the dataset. From there we will be able to determine the most positively and negatively correlated stocks in the market.

Another interesting result is the correlation between opening and closing price. These attributes are strongly positively

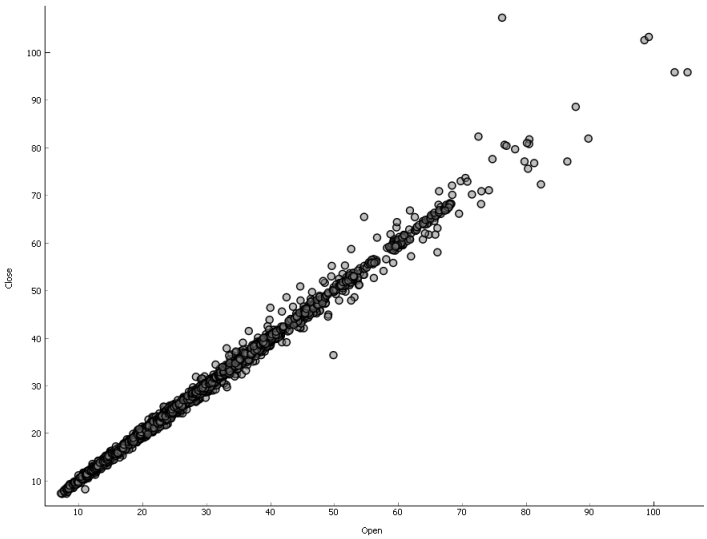

Figure 6: Closing Price vs Opening Price of A

correlated as shown in Figure 6. This result makes sense because on any given day a stock price does not normally change that much. However further mining can be conducted to find stocks with the lowest correlation between opening and closing prices. This would indicate to investors a relatively volatile stock. Some investors prefer volatility while others do not. Knowing this metric, however, helps all investors make a decision about whether or not to add a stock to their portfolio.

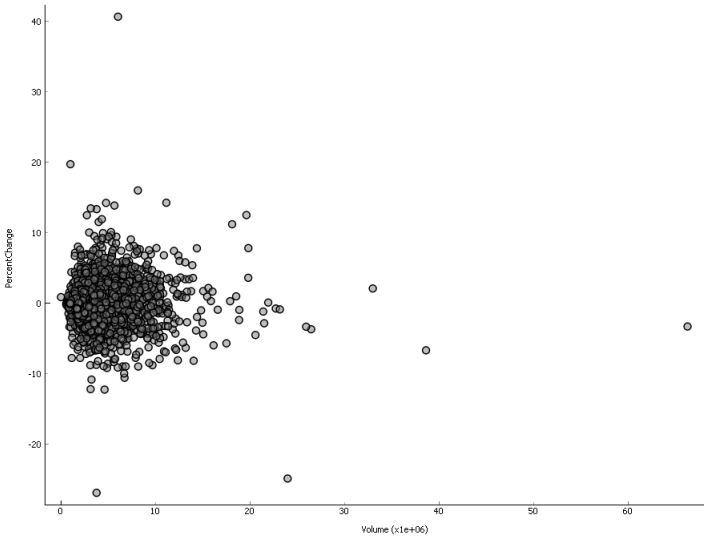

Figure 7: % Change in Price vs Trading Volume of A

The correlation between percentage change in price versus the volume is somewhat unexpected and is shown in Figure 7. High volumes indicates that many people traded the stock on that day. If there are many trades being placed, you might expect that there would be large price movement on that day as well. However in the case of stock A, some of the largest price changes occurred on the lowest volume days. Likewise the higher volume days produced very little change in price. For the next milestone, this comparison will be extended to include many more different stocks in order to gain a better understanding of the prevalence of this pattern. If this pattern is prevalent across most of the stocks, this may indicate that an alert can be generated on low volume days to signal a user there is a buying opportunity.