CSCI 4502 Group 4: Stock Market Analysis - Part 2

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 2
.

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 SP 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)