Project Phase 2 Report

Group 1

1. INTRODUCTION

In this project, we will explore patterns within U.S. stock data through two major data mining techniques: clustering and classification. The stock data is pulled from yfinance, a Yahoo Finance API, and Quandl. Moreover, we are also pulling general economic data, that is yearly GDP and monthly IDP of the U.S., from the World Bank and the Federal Reserve Bank of St Louis. We are expecting if these economic indicators are potential factors affecting the stock and those will be analyzed with the stock dataset through pairwise comparison process. All the data have a time period from 2000 to 2019.

Up until phase 2, we have pulled all the data from our sources and done cleaning on the data-set, including formatting the data, identifying and fixing any problem with the data, and storing them on a MySql database. We have realized the existence of dividends and splits. Simply using open, high, low, and close price of a stock on a specific date, which are directly pulled from our sources, would not give an accuracy result in this project. Therefore, we introduce four extra features: adjusted open, adjusted high, adjusted low and adjusted close price for each instance. These are prices adjusted after dividends and splits happened.

We also have gotten the economic data, which are the annual GDPs and monthly IDPs. However, after doing initial visualizations and pairwise comparisons between them and the stock data, they seem to be irrelevant and we may or may not keep them in the future.

2. DATA DESCRIPTION

This data uses the following elements.

1. `Date:` The date a specific stock trading took place.

2. `Open:` The start price of a specific stock on a specific date. The value of this attribute is pulled from our sources, but it is not used in any of our analysis directly. It is used to calculate Adj. Open, which will be used in our analysis.

3. `High:` The highest price of a specific stock on a specific date. The value of this attribute is pulled from our sources, but it is not used in any of our analysis directly. It is used to calculate Adj. High, which will be used in our analysis.

4. `Low:` The lowest price of a specific stock on a specific date. The value of this attribute is pulled from our sources, but it is not used in any of our analysis directly. It is used to calculate Adj. Low, which will be used in our analysis.

5. `Close:` The final price of a specific stock on a specific date. The value of this attribute is pulled from our sources, but it is not used in any of our analysis directly. The actual attribute we will use is Adj. Close.

6. `Volume:` Total number of stocks traded over the course of a day.

7. `Dividends:` The reward for the stockholders given by the corporation; usually rewarded quarterly or twice a year or monthly, some companies offer a one time dividend across the year. It is usually done when the company is booming with profits. The rewards are offered either in form of cash or some extra stocks(also known as splits).

8. `Company:` The company name of this specific stock. Each instance should have its unique Date and Company value. That is, there is only one stock record per day for each company.

9. `Adjusted closing price:` The adjusted close price of this specific stock on this specific date after a dividend and split happened.These values are calculated using the splits in the stocks.Using these values other adjusted values can be calculated.

10. `Adjusted Open price:` The adjusted open price of this specific stock on this specific date after a dividend and split happened. The data pulled from yfinance does not contain this attribute. Therefore, any instance which has null value on this attribute will be filled up a value through the equation:

$$Adj.Open = \frac{Adj.Close * Open}{Close} \tag{1}$$

11. `Adjusted High price:` The adjusted high price of this specific stock on this specific date after a dividend and split happened. The data pulled from yfinance does not contain this attribute. Therefore, any instance which has null value on this attribute will be filled up a value through the equation:

$$Adj.High = \frac{Adj.Close * High}{Close} \tag{2}$$

12. `Adjusted Low price:` The adjusted low price of this specific stock on this specific date after a dividend and split happened. The data pulled from yfinance does not contain this attribute. Therefore, any instance which has null value on this attribute will be filled up a value through the equation:

$$Adj.Low = \frac{Adj.Close * Low}{Close} \quad (3)$$

13. `Splits:` A ratio indicates how many more stock a stockholder now has after a split happens. It is a rare occurrence and 1 is a placeholder of this attribute indicates there is no split happened on that date. In order to combat over inflated stock prices a company will sometimes perform a split. This means that it will (usually) decrease the price of the stock in exchange for adding more of it into circulation.

14. `GDP:` Gross domestic product: the total value of goods produced and services provided in a country during one year in US Dollars. Simple way of evaluating if this particular trade occurred during a net positive or net negative year, can be used alongside aggregated stock data to show how reliant stock price is on the economy at large.

15. `Industrial Output Index:` A percent production of Industrial production is a measure of output of manufacturing based industries, including those producing goods for consumers and businesses based on the output of the same for 2012.Correlation (or lack thereof) can show how reliant a given company is on demand for US manufactured goods. Because the US has changed to being a service industry over time this is an important statistic as it can indicate the viability of this business into the going future.

3. CURRENT APPROACHES

3.1 Classification

For the classification, since we are not having a nominal target attribute in the data. So we tried to focus on prediction of the stock based on the historic data. This is done by choosing the prior 30 day adjusted closing price as input and 31st adjusted closing price as output. So we get our training data using a sliding window approach. Currently we have chosen linear regression as our machine learning model. The model seems to work good with stable stocks, such as Apple Inc. Unfortunately the model fails predicting highly volatile stocks like Amazon and Alphabet .inc. Therefore, we are still developing the model, trying to generate a general model that works under all circumstances.

3.2 Clustering

To cluster the data, we are planning to work on the Adj. Close attribute with own developed program, as this gives us the closing price of the stock everyday. For each of the 49 companies, we have computed mean, standard deviation, maximum value and minimum value for the adjusted close attribute. Then, with these values, we are planning on clustering the data with two goals :

1. Establish clusters of companies based on their prices (Low, Medium, High and very High);
2. Establish clusters of companies based on how volatile and how fluctuating the stock is. This can be done by using the standard deviation attribute as it shows how much the closing values vary from the mean of the closing values. We intend to use K-Means to perform the clustering, and determine the value of k by using the elbow plot.

3.3 Visualizations

Since all of our attributes are numerical values and they are separated by dates and companies, it is easy to make visualizations of most of them. We are planning to visualize the Open, Close, High, and Low based on companies. We are expecting to see similar patterns and paths each company may have on their own stock price changes. The tools we are going to use for doing the visualizations are WEKA and Microsoft Excel.

3.4 Pairwise Comparison

Since we have the GDP and IDP of the U.S. of each year, we are planning to do pairwise comparison between any of them with average traded stock volume of each year. We are expecting a positive linear relationship in this comparison, that is, if GDP or IDP increases, the annual average traded stock volume should also increase. The average monthly IDP is calculated as:

$$avgIDP = \frac{sumofeachmonthlyIDPinaspecificyear}{12} \quad (4)$$

The average annual stock transfer volume is calculated as:

$$avgVol = \frac{sumofvolumeofallstocksinaspecificyear}{numberofstockrecordinthatyear} \quad (5)$$

A initial pairwise comparison plot matrix of these data are shown in Figure 1:

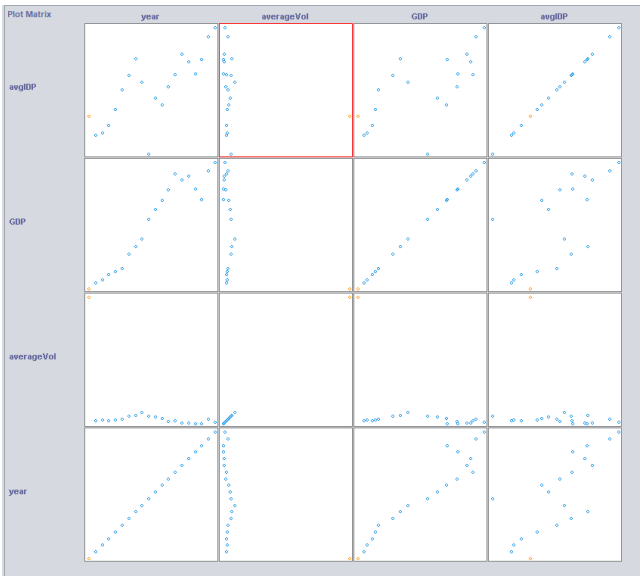

Figure 1: Initial Plot Matrix Generated from Economic Data

Meanwhile, we are also planning to do pairwise comparison between Adj. Open and Adj. Close. We are expecting

to see a positive linear relationship as well since literally, a higher start price should end with a higher end price. The tool we are going to use for the pairwise comparison is Weka. A initial pairwise comparison plot matrix of these data are shown in Figure 2:

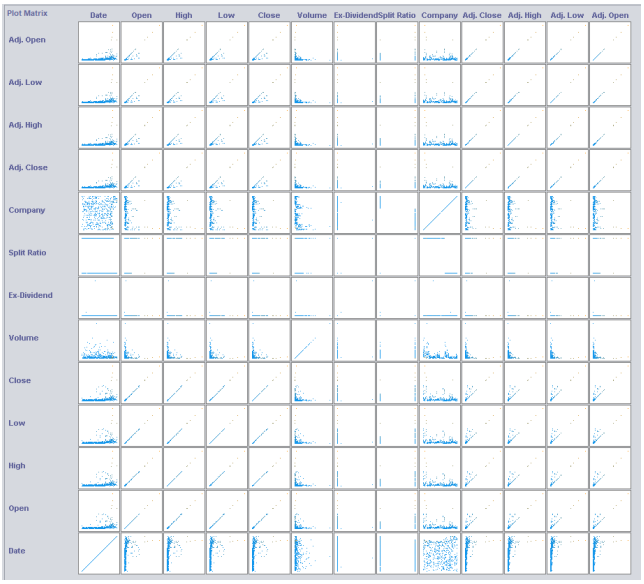

Figure 2: Initial Plot Matrix Generated from Stock Data

4. PLANNED DELIVERABLES

- Develop an accurate model using machine learning techniques for classification .
- Finding distinct and meaningful clusters that help show a relation between the stock prices of the 49 companies.Trying out different clustering algorithms and combination of attributes to identify which method gives the best results.
- Keep working on the analysis on the datasets, while working on the research report, presentation, document all the source files, and README file. Submit all of them as Phase 3 as required at the end of week 10. As well as the peer evaluation.(7/27/2020)