U.S. Stock Analysis

Group 1
July 27, 2020

Introduction

In this project, three data mining techniques: clustering, classification and association mining, are using to explore the patterns of the dataset. Furthermore, four attributes are visualized, and two attribute pairs are pairwise compared to examine any linear association between them.

The stock market is very volatile and goes up and down based on various reasons. Our goal for this project was to analyze the data from the year 2000 onwards to highlight some hidden patterns in the data so that this information could be used to make better trading decisions, and thus maximise the profit gained from investing.

Outlines

- Project Design
- Implementations
- Results
- Lessons Learned
- Current Status and Future Work

Design

Project Setup

The stock data is pulled from yfinance, a Yahoo Finance API, and Quandl. Moreover, we also pulled general economic data, the yearly GDP and monthly IDP of the U.S., from the World Bank and the Federal Reserve Bank of St Louis. All the data have a period from 2000 to 2019 and have been cleaned and stored on a MySQL database.

We have realized the existence of dividends and splits. Only 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 accurate 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.

Dataset Description

- **Date**: The date a specific stock trading took place.
- **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.
- **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.
- **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.
- **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.

Dataset Description (cont.)

- **Volume**: Total number of stocks traded over the course of a day.
- **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).
- **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.
- **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.

Dataset Description (cont.)

- **Adjusted Close**: 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.
- **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.
- **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.

Dataset Description (cont.)

- **Adjusted Open**: 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=(adj. close * open)/close**
- **Adjusted High**: 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=(adj. close * high)/close**
- **Adjusted Low**: 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=(adj. close * low)/close**

Implementations

Classification

Data Preparation:
In this case the prior 30 day stock prices would act as the features to predict the 31’st day stock price.
So we use a basic sliding window approach to prepare our data.

This data is then split into training data and testing data. The 80% of the data is used for training and the other 20% is used for testing.

Training the model:

The model is trained by doing simple linear regression of the training data. Where the input vector is the 30 day price vector and output is the 31st day stock price. The model is trained separately for

Evaluating the model:

We use RMSE to estimate the model. For each company we calculate the RMSE separately and plot it on a bar chart.

Clustering

Two sets of clusters are made:

1. Clustering based on mean closing prices per company
2. Clustering based on variance of closing prices per company

1. The goal of the first clustering is to group similarly priced stocks into clusters so an investor can decide which stocks are similar in terms of their closing price
2. The goal of the second clustering is to group similarly volatile stocks in clusters by grouping on their variance. These clusters show how much a group of stocks vary in their prices over the years.

Association Rule Mining

The goal of generating rules is to see if rules can be generated that gives proof that certain stocks rise together and there is some relation in between them

For the year 2008, each day's data was converted into a format where if the stock price for a specific company was higher on the day than the previous day, it is then included in the dataset.
For example, a row in the data may look like :
Day 1 {'Apple', 'Microsoft', 'Amazon'}

This indicates that between day 0 and day 1, the stocks of Apple, Microsoft and Amazon have risen in price.

This data is then fed into an apriori algorithm with support = 0.3 and confidence = 0.9. With confidence = 0.9, this implies that any rules generated will appear in the dataset 90% of the time.

Visualizations of Attributes

4 attributes are visualized;
Open, Close, High, and Low

Based on 4 randomly chosen companies:

Illumina Inc, Biogen Inc, Exelon Corp, and Ross Stores Inc.

Tool: Microsoft Excel

We expected to see similar patterns each company may have with its stock price path among these four attributes.

Pairwise Comparison

1st pair: Adj. Open and Adj. Close

2nd pair: GDP and Average Annual Stock Transfer Volume.

Method: linear regression with 95% confidence interval

Average Annual Stock Transfer Volume is calculated as: Avg. Vol=sum of vol of all stocks in a specific year/number of stock record in that year

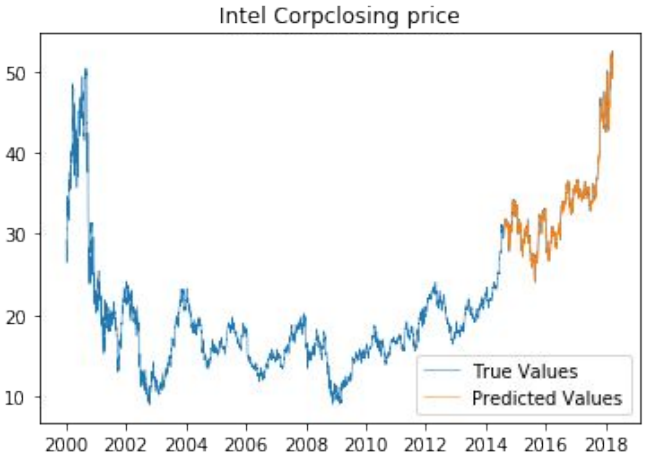
Tools: Weka and Microsoft Excel

We expected to see both pairs having positive linear relations.

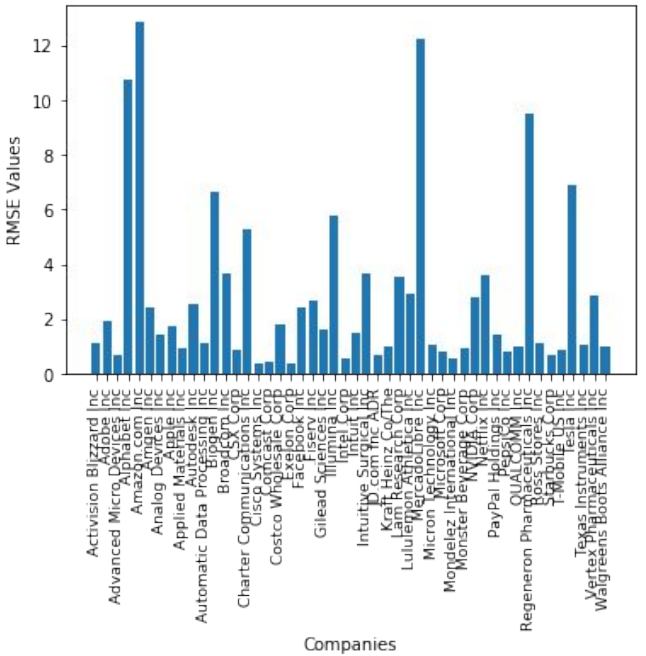
Results

Classification

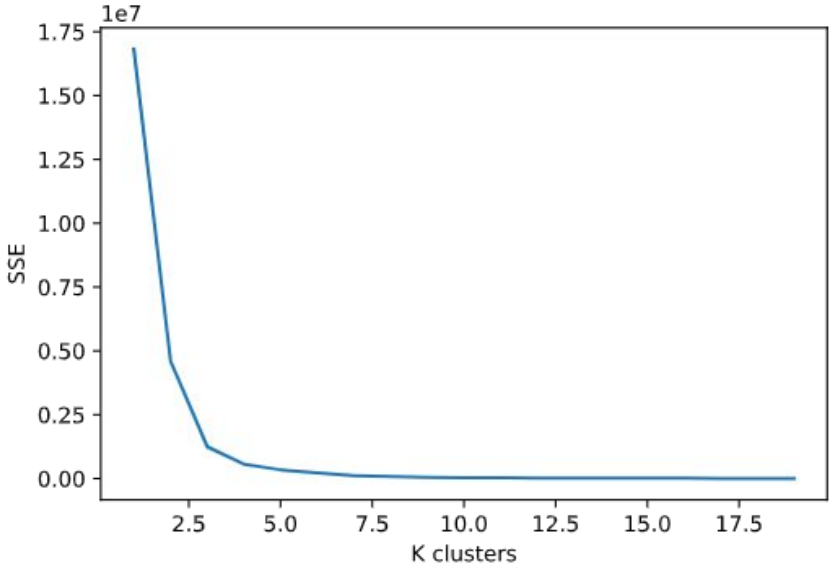
The predictions made by the linear regression can be visualised for a particular stock.



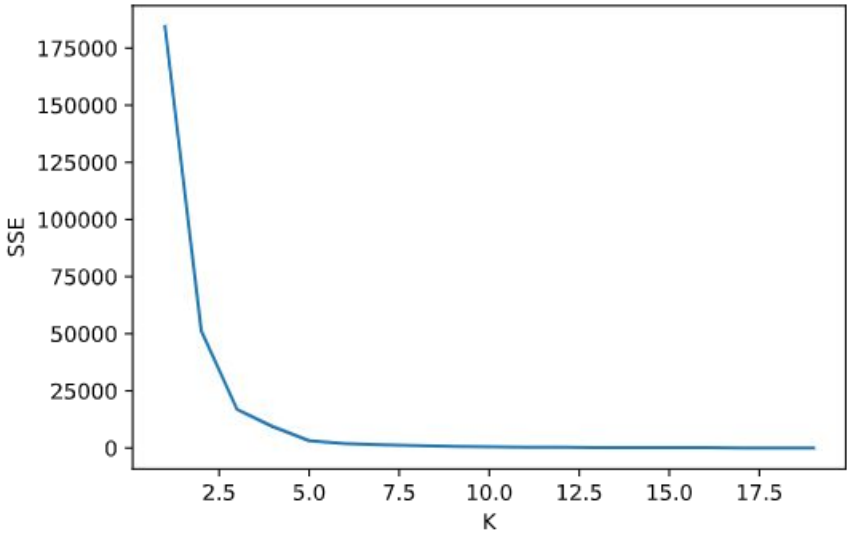
We use linear regression to predict the stock prices. The Rmse is calculated for the predictions of each stock.



Clustering



Clusters A : Knee plot for clustering companies on mean closing price



Clusters B : Knee plot for clustering companies on variation of price

Clustering (cont.)

Cluster	Mean Adj. Price	StdDev Adj. Price	Minimum Adj. Price	Maximum Adj.Price
0	108.874	101.262	9.264	440.913
1	2149.142	37.526	2077.56	2206.09
2	37.077	24.398	9.706	115.979
3	354.751	285.693	28.064	1392.975

Table A : Cluster Descriptions for clustering companies on mean closing price

Cluster	Mean Adj. Price	StdDev Adj. Price
0	182.576	43.112
1	148.894	168.933
2	121.969	100.465
3	354.751	285.693
4	30.158	16.299

Table B : Cluster Descriptions for clustering companies on mean closing price

For Clustering A : Knee-plot shows the best number of clusters was 4. The results of the clustering show 4 distinct clusters.
For Clustering B : Knee-plot shows the best number of clusters was 5. The results of the clustering show 5 distinct clusters.

Association Rule Mining

There were 6 rules generated by the algorithm with support = 0.3 and confidence = 0.9

- {Cisco Systems Inc, Comcast Corp} -> {Automatic Data Processing Inc}
- {Comcast Corp, Costco Wholesale Corp} -> {Automatic Data Processing Inc}
- {Comcast Corp, Fiserv Inc} -> {Automatic Data Processing Inc}
- {Intel Corp, Lam Research Corp} -> {Adobe Inc}
- {Analog Devices Inc, Intuit Inc} -> {Intel Corp}
- {Analog Devices Inc, Lam Research Corp} -> {Intel Corp}

These rules confirm our initial hypothesis that there are some groups of stocks that cause other stocks to rise as they rise.

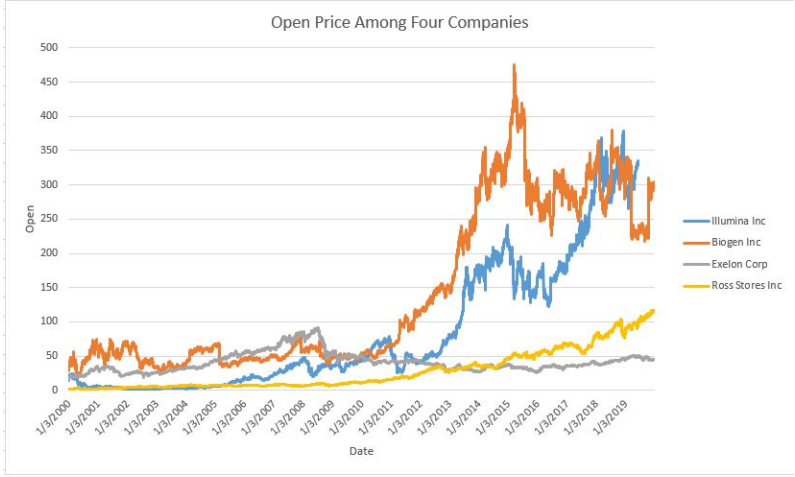
Visualizations of Attributes

The visualizations of Open, High, Low, and Close prices among the four randomly chosen companies (Illumina Inc, Biogen Inc, Exelcon Corp, and Ross Stores Inc) are shown on the next slide:


Low Price Among Four Companies
Low
Date
Illumina Inc
Biogen Inc
Exelon Corp
Ross Stores Inc

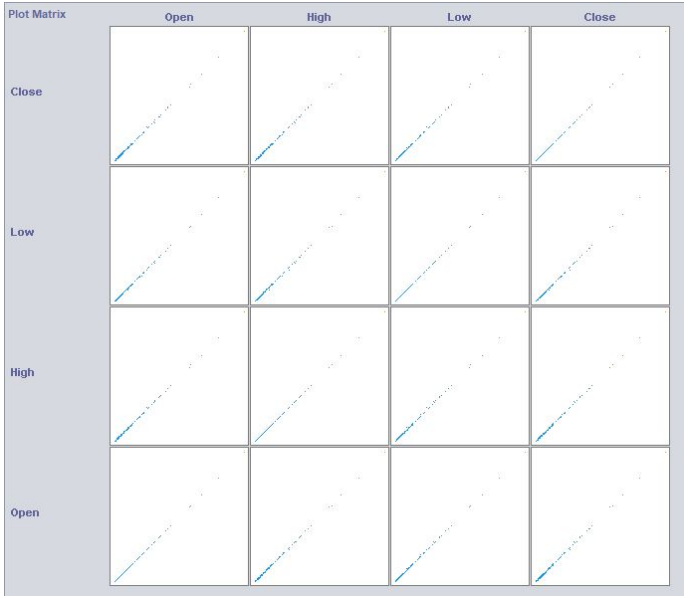

High Price Among Four Companies
High
Date
Illumina Inc
Biogen Inc
Exelon Corp
Ross Stores Inc


Close Price Among Four Companies
Close
Date
Illumina Inc
Biogen Inc
Exelon Corp
Ross Stores Inc

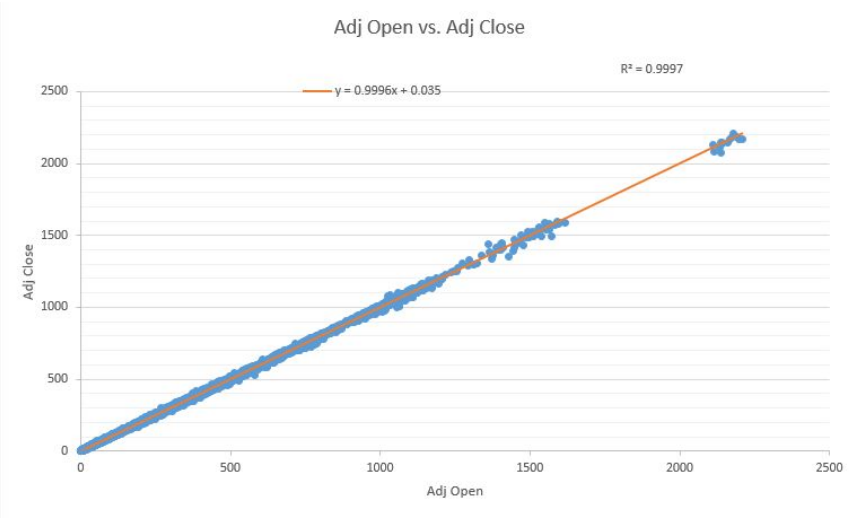

Open Price Among Four Companies
Open
Date
Illumina Inc
Biogen Inc
Exelon Corp
Ross Stores Inc

Visualizations of Attributes (cont.)

As we expected, all four attributes have shown a same pattern for each company, which is not surprising because these four attributes are strongly positively correlated with each other.



Pairwise Comparison Between Adj. Open and Adj. Close



Regression Statistics	
Multiple R	0.999873612
R^2	0.999747239
Adjusted R^2	0.999747237
Standard Error	2.130268867
Observations	119935

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2152721249	2152721249	474371901.7	0
Residual	119933	544261.4047	4.538045448		
Total	119934	2153265511			

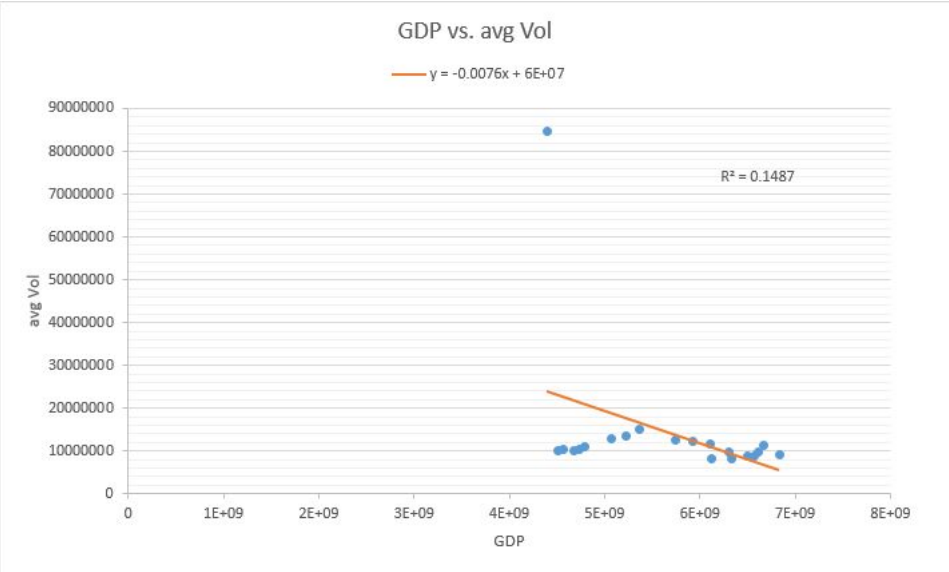
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.034960817	0.007021505	4.979106213	6.3967E-07	0.021198782	0.048722852
x Variable	0.999559099	4.58933E-05	21780.08039	0	0.999469149	0.999649049

Pairwise Comparison Between Adj. Open and Adj. Close (cont.)

The regression equation of the linear relation between adjusted open and adjusted close is y=0.9996x+0.035, with a correlation coefficient of 0.9997. The standard error is 2.1303. With a 95% confidence interval, the lower intercept is only 0.0212, and the upper intercept is only 0.04872. Since the slope of this equation is very close to 1, and the correlation coefficient is very close to 1, we can say that there is a perfect positive linear relation between them.As the adjusted open price of a particular stock increased, the adjusted close price also increased.

We have predicted this outcome because in general, if a stock has a high open price on a particular day, its close price should also be high. This comparison has provided this argument.

Pairwise Comparison Between GDP and Avg. Vol



Regression Statistics	
Multiple R	0.385658
R^2	0.148732
Adjusted R^2	0.101439
Standard Error	15777036
Observations	20

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7.82819E+14	7.82819E+14	3.144924805	0.093083706
Residual	18	4.48047E+15	2.48915E+14		
Total	19	5.26329E+15			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	57425691	24500769	2.343832	0.030765	5951484	1.09E+08
x Variable	-0.00761	0.004289	-1.77339	0.093084	-0.01662	0.001405

Pairwise Comparison Between GDP and Avg. Vol (cont.)

The regression equation of the linear relation between annual GDP and average annual stock transfer volume is y=-0.0076x+6E+07, with a correlation coefficient of 0.1487. The standard error is 15777036. With a 95% confidence interval, the lower intercept is 5951484, and the upper intercept is 1.09E+08. Since the slope of this equation is very close to 0, and the correlation coefficient is very close to 0.1, we can say that there no linear relationship between them, which has surprised us.

We were expecting to see if the economic indicators are potential factors affecting the stock during the design phase. We assumed that higher economic values, i.e., a better social economy, will lead to a higher stock transaction. However, the outcome of this comparison has shown that there is no such relation, at least not linear relation, between GDP and stock transaction volume.

Lessons Learned

Although clustering does not give us any concrete information about which stocks to invest in, it does group stocks into clusters that can help us decide what cluster of stocks we can invest in, and it can also show which category a stock can belong to.

The results from association rule mining show that a pattern can be established to show stocks that rise together. This can help keep track of rising stocks and can help make informed decisions on which stocks to invest in.

The results of linear regression are promising however the predictions are not exact but close to the actual predictions. Certainly these results can be used in real time trading. The same technique can be used to predict the stock behaviour of future month or even a year.

From our pairwise comparison analysis, there is no linear relation between economic indicators and stock transactions. Thus, different from what we expected, economics is not a factor affects stock market.

Current Status and Future Work

Currently the association rules generated are only for one year and only for the rise in stocks. Future work could include our entire dataset for more concrete rules, and also include rules that describe stocks that fall together in price. This would then give further information of which stocks to invest in so that one could make maximum profit.

Moreover, in this project, only the U.S. stock records between 2000 and 2019 are analyzed. To do more reliable and accurate analyses on stock marketing, analyzing stock data from all over the world with a longer time period should be considered in the future.

Currently the stocks are predicted entirely based on historical data. We can add the daily news for a specific company and do the sentiment analysis on the text on each day and it can be considered as a new feature.

Reference

Stock data:

- yfinance https://github.com/ranaroussi/yfinance
- Quandl https://www.quandl.com/

GDP & IDP

- The World Bank Group https://data.worldbank.org/

Thanks