Analysis of S&P 500 from February 2013 to February 2018

By: Jacob Chan, Michael Chyan, Michael Mok

June 5, 2020

CSE 163: Intermediate Data Programming

University of Washington

Summary of Research Questions

1. Which industries in the S&P 500 are growing or shrinking the most over the documented 5 year span?

   We want to research this because it would be interesting, for future projects, to line up our data analysis with current events that had happened in the 5 year span from February 2018 and see how different industries were affected by current events.

   From 2013-2018 the Energy and Telecommunication services sectors were shrinking the most. The Industrial, Information Technology, and Financial sectors grew the most.

2. How accurately can we predict performance of the stock market which is known for its volatility?

   The stock market has been known to be notoriously difficult to predict, so we wanted to see what would happen if we used a machine learning model to come up with projections.

   From our industry machine learning, we can predict general linear regression but not very accurately.

3. Which day of the week, week of the month, or month of the year do stock prices typically rise or fall the most?

   Since our data includes historical stock prices by day, we have a lot of flexibility for our analysis and can use that to find trends at every level.

   Due to time constraints, we did not analyze this. But we would have used something with datetime to determine which day of the week had most gains from the day prior.

4. Which sectors recover the fastest after significant financial downturns?

   Healthcare was relatively unaffected and consumer staples had the quickest recovery.

Motivation and Background

We wanted to examine this topic because of the economic implications of COVID-19. As the stock market is slowly recovering from the initial downturn, we wanted to know which sectors have historically done well and shown the quickest recovery. We can learn which stocks specifically in the S&P 500 recover the quickest and the sectors they belong to. By examining the fallout of the 2008 financial crisis, we can examine how quickly it took for certain industries to recover. However, we also know to take our findings with a grain of salt because of the difference in industries that were primarily impacted in each case.

Dataset

https://www.kaggle.com/camnugent/sandp500
The first dataset we are using is the historical stock prices from 2013 to 2018. It includes the open, high, low, close prices and volume for the day.

https://datahub.io/core/s-and-p-500-companies-financials/r/1.html
The second dataset we are using is to help us connect the industries and companies within the S&P 500, since our first dataset does not provide that. For our analysis we will mainly look at the top 250 companies to hopefully take account for companies that fall in and out of the S&P 500. (But we are open to ideas to work with this). This data set is from two years ago and we are mainly wanting the columns of industry and name of company to join to our other dataset.

https://www.kaggle.com/qks1lver/amex-nyse-nasdaq-stock-histories?
This dataset includes all historical data (up to 2019) from the NYSE, NASDAQ, and AmEx exchanges. With this information, we have a lot more historical data that can be used to evaluate performance in specific time periods of interest. In our case, we can compare current times with the 2008 housing financial crash.

https://finance.yahoo.com/quote/%5EGSPC/history?period1=1431561600&period2=1589414400&interval=1d&filter=history&frequency=1d
This dataset is for our personal interest because we wanted to extend our machine learning model all the way out to 2020 which is outside the scope of our main dataset. This required us to find another dataset to compare the overall trend of stocks in the S&P 500 to the S&P 500 *index*.

Methodology

There are two main components to our project. The first is that we are aiming to create a machine learning model that can create at least a 70% accurate projection of the stocks in the S&P 500. We chose the S&P 500 since these companies are supposed to be relatively stable as they are the largest companies by value. Since there is too much information available, we want to create a projection primarily for 2013-2018. However, for our own curiosity, we will extend our model past 2018 using other datasets to test accuracy of our model. We would also create a second machine learning model that performs similarly except it would group the stocks by industry. This leads into the second component of our project.

The second part is that we want to do some industry specific research with regards to financial crises. We will be joining several tables together to determine how certain sectors respond to economic downturns. This would also allow us to see which industries are capable of recovering the quickest and returning to original growth projections. By conducting analysis on both individual stocks and sectors, we can note any outliers and also do external research on reasoning behind business decisions during and after these crises. We will be primarily focusing on the 2008 housing crash since it is most similar to our current COVID-19 crisis in terms of abruptness, but we will also be keeping note of the differences between the two scenarios.

We will also be including visualizations for both of these because of how number-heavy and dense the data is. By using visualizations, it makes the output much easier to understand for someone who has never had experience with stocks.

Results

Which industries in the S&P 500 are growing or shrinking the most over the documented 5 year span?

From 2013-2018 the Energy and Telecommunication services sectors were shrinking the most. The Industrial, Information Technology, and Financial sectors grew the most. From our graphs and linear regression lines that we made from the ML models we could easily see the trends and the shape of each graph and how they changed over time. Our ML linear regression model also gave us an understanding of what the trend line was and showed the increase or decrease of each sector.

How accurately can we predict performance of the stock market which is known for its volatility?

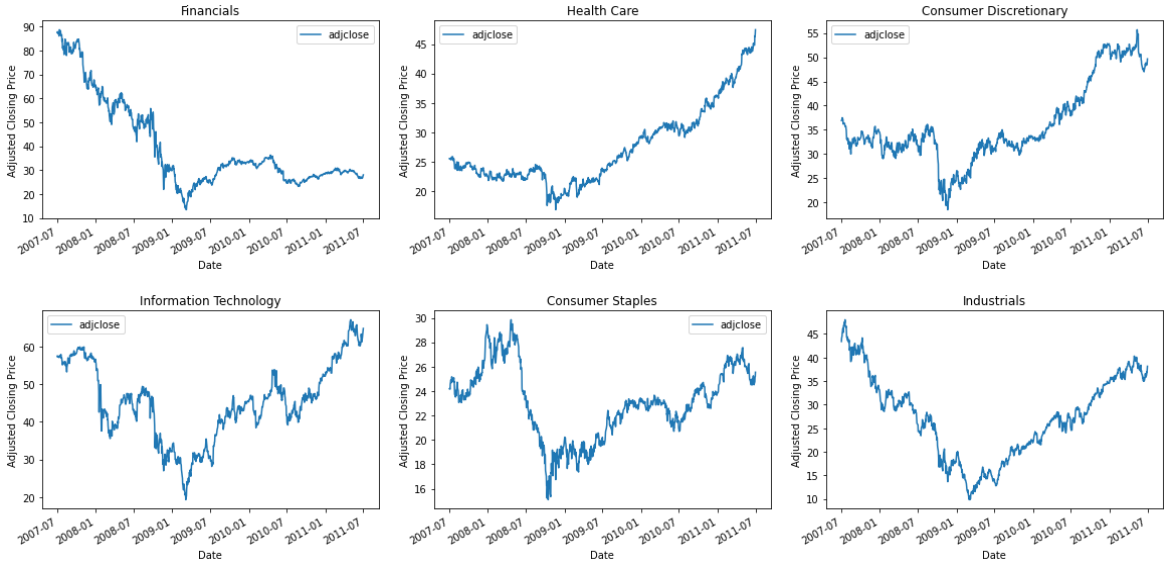
From our sector analysis using ML linear regression models, we can somewhat predict accurately the trend of the lines year to year timelines however our linear regression model cannot predict day to day or month to month timelines. Our use of the Mean Squared Error value also did not provide a lot of accurate understanding because the MSE values that were the lowest had the most volatile and least amount of linear trend for that sector.

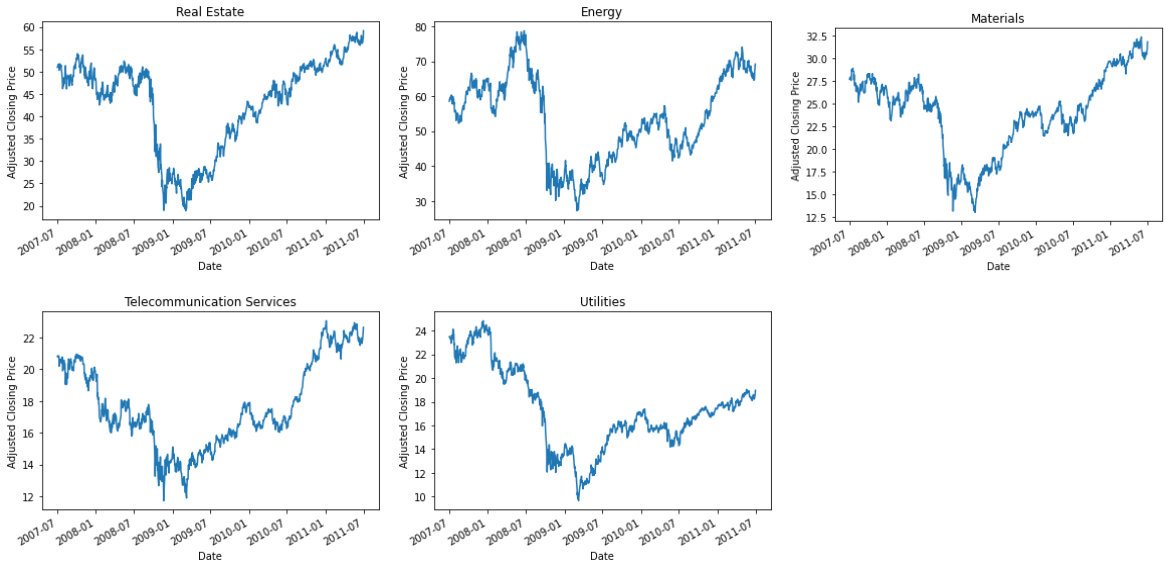
Which day of the week, week of the month, or month of the year do stock prices typically rise or fall the most?

We did not get to this question due to time constraints and the complexity of coming up with a specific criteria to measure "most growth". However, if we had more time, we would have tried something with datetime to extrapolate the day of the week and then use the change from the day prior to decide the greatest average change.

Which sectors recover the fastest after significant financial downturns?

Specifically for the 2008 financial crisis, we found that different industries were impacted very differently (when they were hit, severity of the impact,etc.) so it was hard to say exactly which recovered the fastest. However, Healthcare was relatively unaffected and actually continued to grow steadily. Of the industries that were hit harder, consumer staples took a fairly steep dive, but recovered fairly quickly.





Challenge Goals

Machine Learning

Our project completed our goal of using machine learning models. However we had to diverge from our original goal of having 70% accuracy because our data that we would use did not fit an accuracy score and instead needed regression models to fit. We would have liked to use other machine learning models other than just linear regression but we still accomplished our goal of using machine learning.

Combining Multiple Datasets

Our project completed our goal of combining multiple datasets because we needed to combine two datasets (which we predicted accurately in our challenge goal). For the Machine Learning Model for analyzing industries over time, our team had to combine a S&P 500 data set that had industries and a daily market value of S&P 500 companies that did not have specified industries.

New Library

Our project used plotly.graph_objects to analyze historical data and financial crisis analysis to create better graphs to show historical data over the years.

Work Plan Evaluation

Original Plan

We plan to complete this project using a version control system. Using Git and github will allow us to collaborate on the code. Each of the tasks will be divided between the members to work on. We will have one person combine the datasets, two people work on the machine learning model, and all members work on the financial crisis analysis.

- Manipulate Datasets (3 hour)
    - Join the S&P 500 dataset from kaggle with the industry one from data.io and the kaggle dataset with historical stock information
        - We will be filtering out all companies not in the S&P 500
        - We want the following columns: sector, name of company, ticker symbol, date, daily opening price, daily closing price, daily low, daily high
- Create Machine Learning Model (5 hours)
    - Stock Projection
        - Test accuracy using the dataset obtained from yahoo finance
    - Categorize by Industry
        - Test accuracy using mean price of stocks in certain industry
- Financial Crisis Analysis (7 hours)
    - Use the historical dataset to analyze stock prices from 2006-2010 to provide context to 2008 crash
    - Determine which industries are most robust
    - Possibly project findings onto current market with COVID-19 and compare results
- Writing the report (5-10 hours) on google docs for collaboration
    - We will collaborate and divide our findings to equally share writing the report
    - Call on zoom calls to edit and revise paper to make sure data matches our report findings.

Evaluation

Using a version control system proved to not be as effective for our team. Two of us had the knowledge on how to use git, but it was difficult to learn in the span of the project for the third member. We then turned to Google Colab. This helped us have a space to play around with our datasets and share our code close to real time. Pair programming ended up coming into play for a few of our coding sessions. Often we only had two tasks that needed to be done at a time, which allowed us to dedicate two members towards one goal. This helped increase the efficiency of our programming time. We still decided to work with github as a place to store and organize our scripts and datasets.

In terms of timing for our deliverables the timing was either slightly an understatement or relatively accurate. However we did not accomplish the same deliverables or methodology that we originally decided on. Our machine learning models had to be changed because we did not predict the models we needed to use to represent our data effectively. We also did not make test files or attempt our original “Create Machine Learning Model - Stock Projection - Test accuracy using dataset obtained from yahoo finance” because we ran out of time and it was becoming difficult to work with the code and felt like it was more than we could handle.

We effectively were able to analyze our original work plan category of “Creating Machine Learning Model - Categorized by Industry - Test accuracy using mean price of stocks in certain industries”. We used linear regression machine learning models to analyze by industry, and while not very accurate we still were able to analyze and create a machine learning model. However we were not able to obtain our original goal “The first is that we are aiming to create a machine learning model that can create at least a 70% accurate projection of the stocks in the S&P 500.”, because we did not use models that could test percentages like this.

For the financial crisis analysis, we were able to generate graphs that showed how industries performed before, during, and after the 2008 crash. We used visualizations to come up with our analysis.

Testing

Our team did not make any test files because the code we wrote was made specifically for the data sets we used, e.x hardcoded some of the lengths and we simply ran out of time.

Collaboration

For this project, we did not get help from any students in the class or other people. We, however, often used stackoverflow and Ed’s discussion board to answer any questions we have. We also use the documentation for pandas and other libraries frequently.

Notes

Quarantine has tired and drained so we were unable to put forth our best efforts. Sorry for disappointing. These last few weeks have been extremely difficult and we regret our shortcomings and wish we could’ve done better.

Love you hunter <3
Thanks for a great quarter