Project Executive Summary

Group 2

November 10, 2021

Introduction and Background Information

I. Introduction

Studying and learning from financial trends is a key aspect of financial management. The stock market experiences changes throughout the economic cycle and can be affected by major events. The finance industry is one of the largest industries trading in the stock market. Drawing from a small sample of financial institutions with a range of market capitalization sizes as well as the S&P 500, we use historical stock price data and machine learning to understand and explain the market and its changes throughout the past and potential changes in the future.

II. Exploratory Questions

1. Which subcategory of the finance industry leads in sales/revenue?
2. How do companies perform compared to each other in the stock market? Are there any common characteristics between the top performing stocks?
3. How do major events such as the Great Recession, presidential elections, and Coronavirus pandemic impact stock performance? How do expectations affect these outcomes?
4. How does the finance industry tend to perform in comparison to the S&P 500 in terms of stock price changes?
5. How accurately can we predict whether a stock will perform better or worse in comparison to the S&P 500?

III. Data Sources

We sourced our contextual data from the United States Census Bureau and financial data from *The Wall Street Journal*. The data from *The Wall Street Journal* provides the opening, closing, high, and low prices for the specified companies as well as the S&P 500 along with the volume traded each day from July 2, 2007 until October 27, 2021.

Links to these sources can be found in the References section at the end of this report.

IV. Data Selection

Due to the large scale of the finance industry, we limited our data to six financial institutions with varying market capitalization sizes in an attempt to reflect the diversity of the market. The six companies that we used in our research are as follows:

Company Name	Ticker Symbol	Market Capitalization	Size Classification
JPMorgan Chase & Co.	JPM	$495.83B	Large

Goldman Sachs Group Inc.	GS	$135.71B	Large
Discover Financial Services	DFS	$34.23B	Medium
Synovus Financial Corp.	SNV	$7.17B	Medium
New York Community Bancorp Inc.	NYCB	$5.86B	Small
Bank of Hawaii Corp.	BOH	$3.52B	Small

Our size classification was largely arbitrary, as the classifications change depending on the source, but we have determined that maintaining an even ground with two large, two medium, and two small companies was appropriate for our research.

Discover Financial Services was the latest of the six to begin trading on the New York Stock Exchange, which was on July 2, 2007. As a result, we decided on utilizing the last 14 years of historical stock price data beginning from that date until October 27, 2021. From this data, we are able to observe many major events in the United States and worldwide such as the Great Recession, four presidential elections, and the Coronavirus pandemic.

Results

I. Total Revenue

The census summary statistics dissected the finance industry into multiple sub-categories, and those sub-categories were further split into many smaller divisions. Figure 1 below displays the total revenue of those major categories by their assigned NAICS code. Insurance carriers and related activities (524) leads with approximately $2.34 billion in revenue, followed by credit intermediation and related activities (522) with $0.99 billion, securities, commodity contracts, and other financial investments and related activities (523) with $596 million, and monetary authorities (521), specified as the central bank, with $117 million.

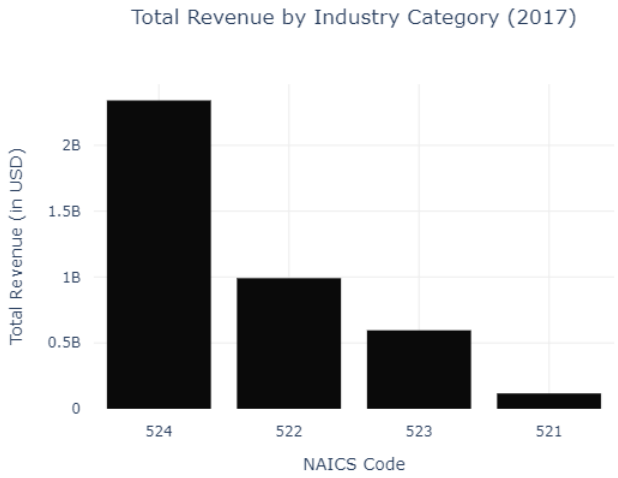

Figure 1. Total Revenue by Industry Category

II. Company Performance in the Stock Market

As the companies for our research were chosen based on their market capitalization, Figure 2 appears to somewhat accurately reflect our categorization of company size in terms of stock prices. While the ranking of companies in our graph has changed over the last 14 years, they generally settle into their expected places by October 2021, except in the case of Synovus Financial Corp. Market capitalization is calculated by multiplying the share price by the number of shares outstanding, so it is to be expected that there is a correlation in large companies having higher stock prices. Larger companies also tend to have a larger presence throughout the country and internationally. In Figure 2, the companies on the lower end of the average closing price by October 2021 are Bank of Hawaii Corp, Synovus Financial Corp., and New York Community Bancorp Inc. The scope of these organizations are generally limited to smaller regions within the United States.

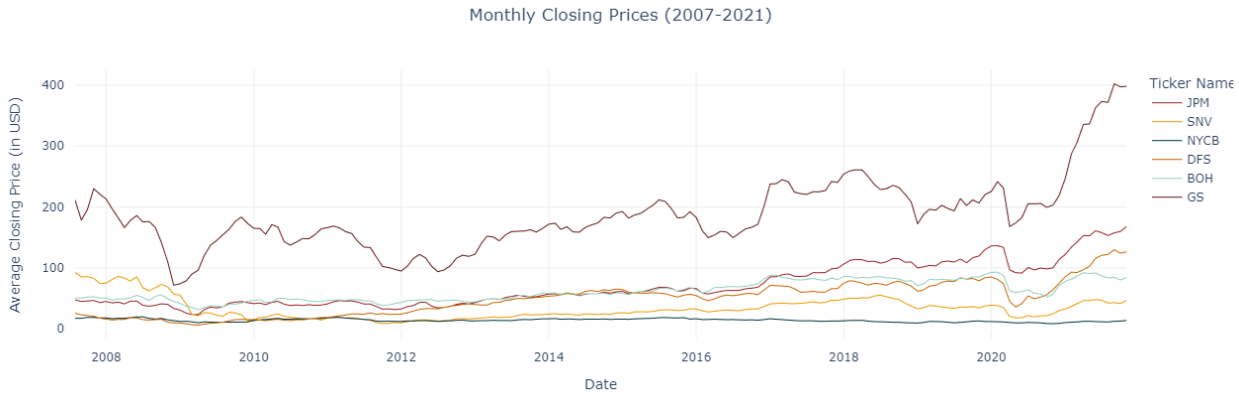

Figure 2. Monthly Closing Prices (2007-2021)

Figure 2 also reflects the impact of major events in the United States and worldwide, with the major drops in prices occurring between 2008 and 2009 as a result of the Great Recession as well as in the spring of 2020 due to the Coronavirus pandemic. While the results are somewhat expected due to the nature of these events, the magnitude can further be quantified through our findings.

Figure 3 exhibits the monthly percent changes in average closing prices for the finance industry compared to the S&P 500 index. The S&P 500 is a measure of the stock performance of the 500 largest companies trading in the United States. Three of the companies used in our research, JP Morgan Chase & Co., Goldman Sachs Group Inc., and Discover Financial Services, are included in the S&P 500. The finance industry, relative to the S&P 500, has experienced a large amount of loss during the Great Recession and a significantly larger amount during the Coronavirus pandemic when governments enacted “stay at home” orders. These results can suggest that the

finance and insurance industry is slightly more reactive to events than other industries represented in the S&P 500.

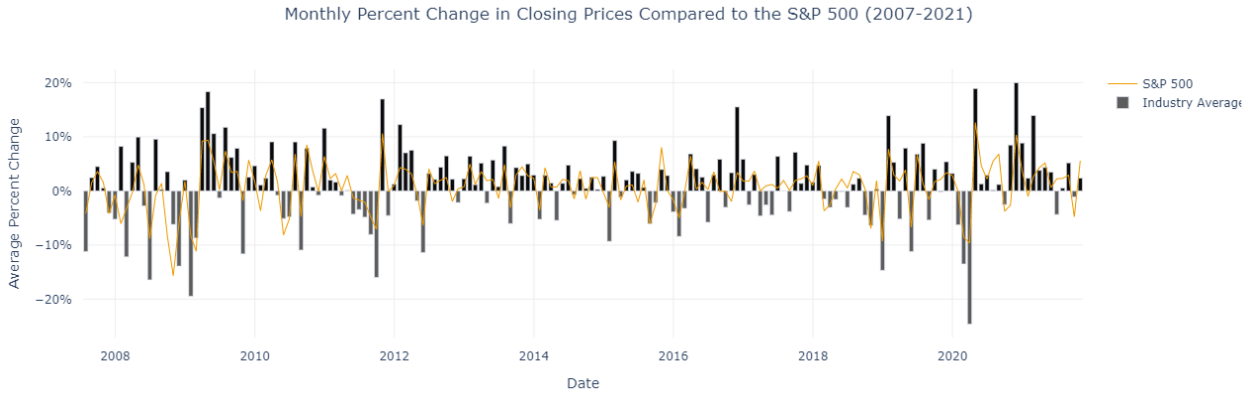

Figure 3. Monthly Percent Change in Closing Prices Compared to the S&P 500 (2007-2021)

In terms of the impact that presidential elections have on the stock market, the results are largely unreliable. The 2008 election occurred during the Great Recession when stock prices were relatively at their lowest. In addition, only one election would be considered an upset in which the predicted winning candidate lost (2016). When comparing the different years, there is no clear pattern as to whether an election's results affected the stock market in a significant way.

However, in Figure 3, the negative values from May to September 2011 unexpectedly became an area that we wanted to explore. The consecutive negative drops can be explained with two possible events. In August 2011, the credit rating for the United States government was downgraded by credit rating agencies, and during the second half of the year, foreign markets were undergoing a debt crisis that impacted the market in the United States.

III. Machine Learning Findings

To address the question that seeks to explore how accurately we can predict whether a stock will perform better or worse than the S&P 500, a logistic regression model was chosen as the approach for the machine learning model.

In the transformation stage of the ETL process, a new data field is created that calculates the percent change of the closing price of one day compared to the day before for both the financial data and the S&P 500 data. A percent change is chosen to help normalize the data as the values among the stocks and the S&P 500 do not allow for a price change to be an accurate measure of change over time. In order to run the logistic regression model, an outcome column is created that is a binary (0 or 1) value that generates a value of 1 when the financial company's stock percent change is greater than the S&P 500's percent change for a given day and a value of 0 when the financial company's stock percent change is not greater than the S&P 500's percent change for a given day.

In addition to that, other changes are made to the data. Rather than make a time series with the date, new columns are added for the year, month, and day. The model itself only retains the Month and Day columns as the inclusion of Year led to poorly performing models.

In order to find a way to include the market capitalization into the model, a new column is created that places each financial stock into a category based on its market capitalization (Large, Mid, Small). That new column is then turned into dummy variables for the model. An exponential weighted moving average of the closing price of the financial stock is created as well. Rather than using a simple moving average which weighs each data point equally, the exponential weighted moving average places more weight on more recent results.

Another aspect to include in the model is the inclusion of using previous days' differences between the stock's percent change and the S&P 500's percent change. In order to find the value, first a new column is created as a reference that was the difference between each financial stock's percent change column and the S&P 500's percent change column. Then the previous day's value can be found and included in the row of data for a given stock on a given day. Figure 4 shows the Bank of Hawaii's first 6 days in the dataset. The column Diff1 takes the previous day's Perc_diff which is the difference between the financial stock's percent change and the S&P 500's percent change. The first value of 0.003498 shifts down by a day for each new column added all the way to Diff5 which takes the Perc_diff value from 5 days before.

Market	Perc_diff	Diff1	Diff2	Diff3	Diff4	Diff5
Small	0.003498	NaN	NaN	NaN	NaN	NaN
Small	-0.001867	0.003498	NaN	NaN	NaN	NaN
Small	-0.004826	-0.001867	0.003498	NaN	NaN	NaN
Small	-0.017499	-0.004826	-0.001867	0.003498	NaN	NaN
Small	-0.007903	-0.017499	-0.004826	-0.001867	0.003498	NaN
Small	-0.003542	-0.007903	-0.017499	-0.004826	-0.001867	0.003498

Figure 4. Table with Added Difference Columns

The additions of these columns are meant to find more features to include in the model while avoiding the inclusion of highly correlated values, as a stock's open, high, low, and close values are all highly correlated.

The chosen model gives a score rating of 0.819, with a Matthews Correlation Coefficient of 0.638. That model includes the percent change of the financial stock, the month of the year, the day of the month, the exponential weighted moving average of the closing price, the differences

of the previous 5 days percent difference between the financial stock and the S&P 500, and the dummy variables for the market capitalization of each stock.

Figure 5 below visualizes the plotting of the data using the percent change of closing price as the value on the x-axis. The y-axis is the binary output column that describes if the percent change of a financial stock's closing price is greater than the percent change of the S&P 500's closing price on the same day. The plot visualizes how, in a general sense, the increase in the percent change from the previous day, increases the probability that the financial stock will outperform the S&P 500 for the day. With a lot of overlap in percent changes near zero.

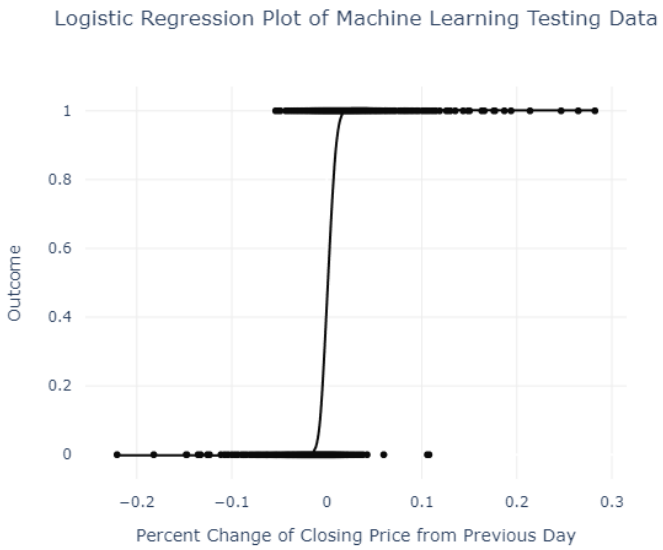

Figure 5. Logistic Regression Plot of Machine Learning Testing Data

Figure 6 displays the Receiver Operating Characteristic (ROC) curve. The area under the curve below matches the score value found from the model at 0.819. With a value over 0.8 this model can be interpreted as a good model with this data.

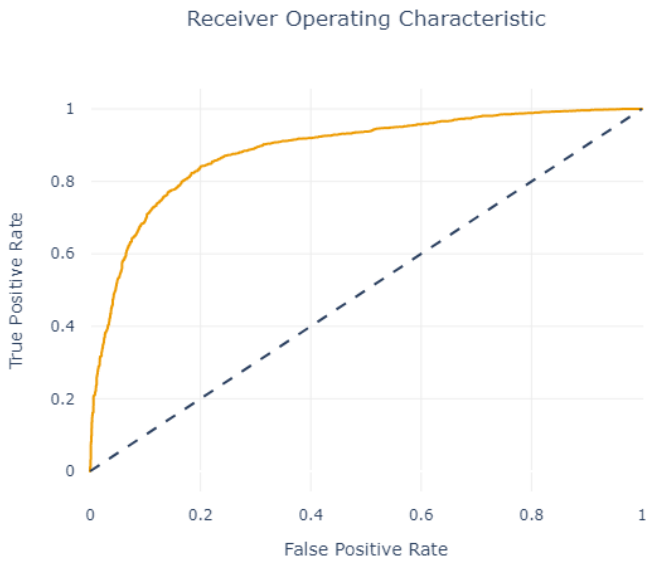

Figure 6. Receiver Operating Characteristic

Due to the model's score, 0.819, and Matthew's Correlation Coefficient 0.638, these results suggest that the model is able to accurately predict the outcome. Using the data adjustments presented above, we are able to predict if a stock's percent change will outperform the percent change of the S&P 500, with a score value of approximately 0.82.

Conclusion and Recommendations

Through our research, we sought to answer multiple questions pertaining to the finance industry and its performance in the stock market. Our findings showed that insurance carriers generate the largest share of revenue within the industry, major events typically follow expected results in the market (except in the case of presidential elections where it can not be proven to have a major impact), and the industry tends to trend similarly to the S&P 500, though with larger drops in prices. The results suggest that the industry is performing well financially in its current state, especially in the context of the recent (and ongoing) pandemic.

In terms of the machine learning model, with more time and more access to historical data, the inclusion of financial data (such as financial statements) would present a more in-depth approach to the machine learning model our group is interested in analyzing. The main issue we ran into is the lack of available data that goes beyond the last 5 years and the nature of financial data being released quarterly creates the issue of only 4 data points per year. The ability to include a company's profit share or debt value adds important components to a company's stock price and it would be interesting to see how it affects the percent change of the stock price from one quarter to the next.

References

The Wall Street Journal. (n.d.) *Market Data.* Retrieved from: https://www.wsj.com/market-data

United States Census Bureau. (2017). *Finance and Insurance: Summary Statistics for the US, States, and Selected Geographies: 2017.* Retrieved from: https://data.census.gov/cedsci/table?q=EC1752BASIC&tid=ECNBASIC2017.EC1752BASIC