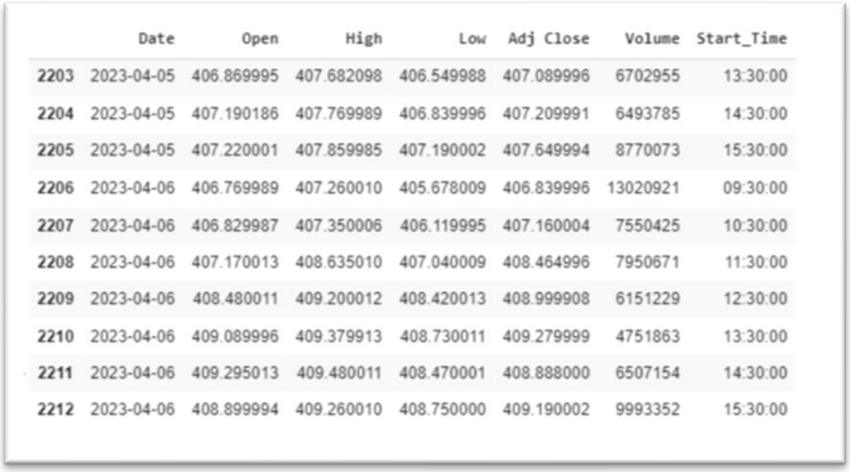
Group 6- Big Data Ultra
Maria Greenfield Brandon Diaz-Lopez Brendon Abbott Jeremy Bouhadana

Final Project

The stock market can seem like an extremely risky and unpredictable environment in which to place one's hard-earned money. There are numerous factors that come into play when it comes to a stock's end-of-day worth, and every investment comes with some amount of risk. Taking all the contributing factors into consideration to decide which stock to invest in can be very overwhelming for one person, and typically one will hire a team of financial advisors to do so for them. Machine learning, however, has developed enough to subsidize a lot of the work involving large quantities of data, and can read through and perform analyses on it in a fraction of the time a team of people could. For this reason, our group has chosen to utilize machine learning to tackle the issue of accurately predicting stock market prices.

As previously stated, there are many contributors that can affect a stock's performance, for example, politics, economics, news, sentiment, popular culture, and natural disasters to name a few. For the purposes of this project, we have chosen to limit our scope to only basing our predictions off the date, time, volume of shares and stock price, more specifically the open, high, low, and adjusted closing price of the stock. We also have split our analysis into three separate datasets that break down the stock's prices on a weekly, daily, and hourly time frame. **Figure 1** below shows an example of the contents of our datasets. The source of our data comes from the S&P 500, which is a Stock Market Index for the largest publicly traded companies in the US. It is essentially a measure of performance for a particular group of companies. The scope of our project will encapsulate the data from a time period staring in January 2022 and ending in the first week of April 2023.



	Date	Open	High	Low	Adj Close	Volume	Start_Time
2203	2023-04-05	406.869995	407.682098	406.549988	407.089996	6702955	13:30:00
2204	2023-04-05	407.190186	407.769989	406.839996	407.209991	6493785	14:30:00
2205	2023-04-05	407.220001	407.859985	407.190002	407.649994	8770073	15:30:00
2206	2023-04-06	406.769989	407.260010	405.678009	406.839996	13020921	09:30:00
2207	2023-04-06	406.829987	407.350006	406.119995	407.160004	7550425	10:30:00
2208	2023-04-06	407.170013	408.635010	407.040009	408.464996	7950671	11:30:00
2209	2023-04-06	408.480011	409.200012	408.420013	408.999908	6151229	12:30:00
2210	2023-04-06	409.089996	409.379913	408.730011	409.279999	4751863	13:30:00
2211	2023-04-06	409.295013	409.480011	408.470001	408.888000	6507154	14:30:00
2212	2023-04-06	408.899994	409.260010	408.750000	409.190002	9993352	15:30:00

Figure 1: Top 10 rows of hourly data from S&P 500

In order to better understand our data, we must implement some Exploratory Data Analysis (EDA's). First, we ran some descriptive statistics on our data and found an interesting detail that across all three of the datasets. The stock prices in each dataset were very different, but the standard deviations of were all within one unit of each other, meaning that they are all similarly distributed. This implies that whatever models are chosen for this analysis, they should work similarly across the three datasets. Additionally, we implemented a series of plots for each of the datasets to bet a more visual understanding of this behavior. The three plots chosen were scatterplots, boxplots, and histograms. The scatterplots seemed very random and did not provide much information (**Figure 2**). The boxplots, however, all had symmetrical distribution and showed no outliers (**Figure 3**). The histograms also had some significant features (**Figure 4**). The Daily and Hourly data displayed some unimodality and right skewness, but shape of the Weekly data was harder to discern.

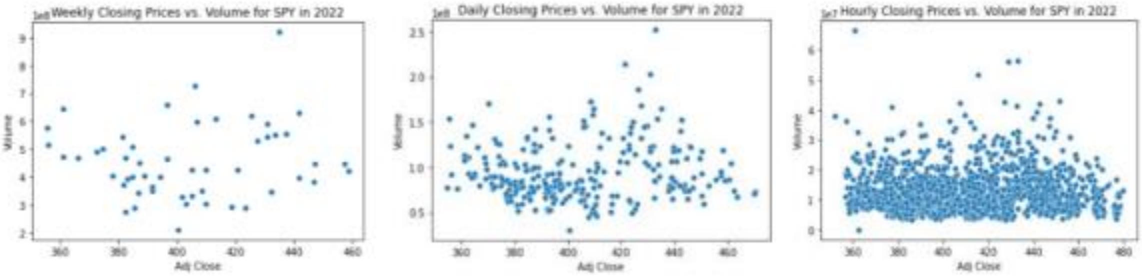


Figure 2: Scatterplots of Weekly, Daily, and Hourly data (from left to right)

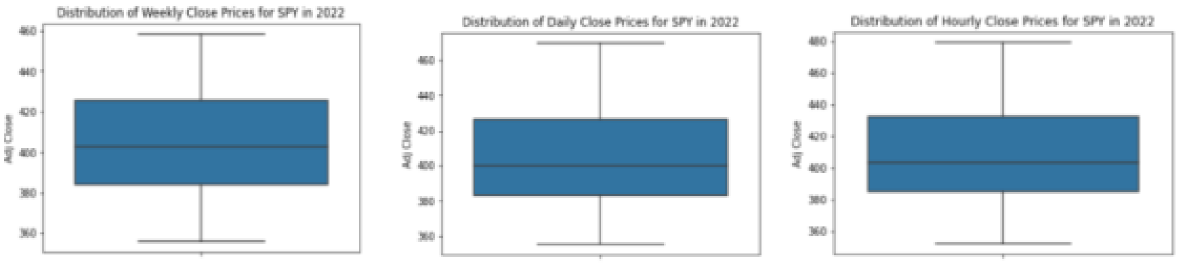


Figure 3: Boxplots of Weekly, Daily, and Hourly data (from left to right)

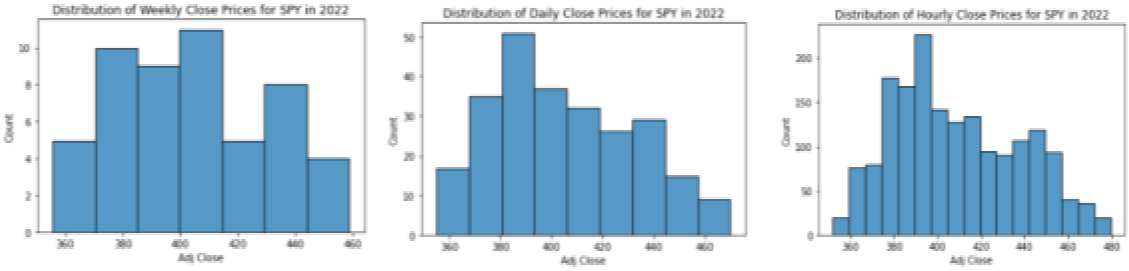


Figure 4: Histograms of Weekly, Daily, and Hourly data (from left to right)

For the matter of predicting the market as accurately as possible, we decided to test the following four different models: Support Vector Machines – Regressor method, Linear Regression – Gradient Descent method, Random Forest Regressor, Linear Regression – Normal Equation method. Our goal is to determine what the adjusted closing price of a stock price will be based on the inputs of ‘Open’, ‘Low’, ‘High’, and ‘Volume’. However, before testing these methods, we must do some preprocessing on the data beforehand. First, since these are time series datasets, we must check for trends within the data. We utilized the Augmented Dickey-

Fuller (adfuller) test to detect any trends in the datasets. The only dataset that displayed non-stationarity was the Weekly, dataset. Detrending data is important because it can allow us to see any other potential sub trends. We can detrend this data by fitting it to a regression model and calculating the difference between the observed and predicted values of the model, otherwise known as the accuracy. Another prepossessing step we must implement is converting our time series dataset into a supervised learning dataset, which we did through the use of a function called 'series_to_supervised()'. Throughout each of our models, we also implemented the MinMaxScaler function, which helps to avoid any bias towards larger values.

The first model chosen, which will be our baseline for the other models, is the Support Vector Machines – Regressor Method. For this model, the best metrics for success are the Mean Squared Error (MSE) and the Coefficient of Determination (R^2). The MSE is a measure of the sum of the variance of the estimator and the squared bias of the estimator. The R^2 value is a statistical measure of how differences in one variable can be explained by the differences in a second variable. The Weekly data yielded an MSE of 0.41 and an R^2 of 0.55, Daily data yielded an MSE of 0.06 and an R^2 of 0.94, and Hourly data yielded an MSE of 0.01 and an R^2 of 0.99. Weekly data had the work results, and Daily and Hourly data had good results. However, Hourly data had a very high accuracy, which indicates the possibility of overfitting.

The next algorithm we tested is the Linear Regression – Gradient Descent method. This machine learning model iteratively updates the parameters based on a specified learning rate. The best metrics for this model are the MSE and the R^2 values as well. Initially we had not transformed the data to a supervised learning dataset, which resulted in very random results. After transforming the data, however, the Weekly data resulted in an MSE of 0.4965 and an R^2 of 0.4568, the Daily data resulted in an MSE of 0.0741 and an R^2 of 0.9258, and the Hourly data resulted in an MSE of 0.0082 and an R^2 of 0.9917. These results are a lot more consistent with what we understand about the data.

The third model tested for this project was the Random Forest Regressor. This machine learning model can handle datasets that contain continuous variables and can provide a higher level of accuracy in prediction over other regression algorithms. The metrics for success used for this model were the Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2). The RMSE is the standard deviation of the residuals or 'prediction errors'. Weekly data yielded an RMSE of 0.1709 and an R^2 of -0.0331, Daily data yielded an RMSE of 0.0568 and an R^2 of 0.9332, and Hourly data yielded an RMSE of 0.0168 and an R^2 of 0.9932. Once again, the weekly data performed the worst, daily performed the best, and hourly data was the closest to potential overfitting. **Figure 5** below shows regression plots of each of the datasets and how closely the models fitted to the data.

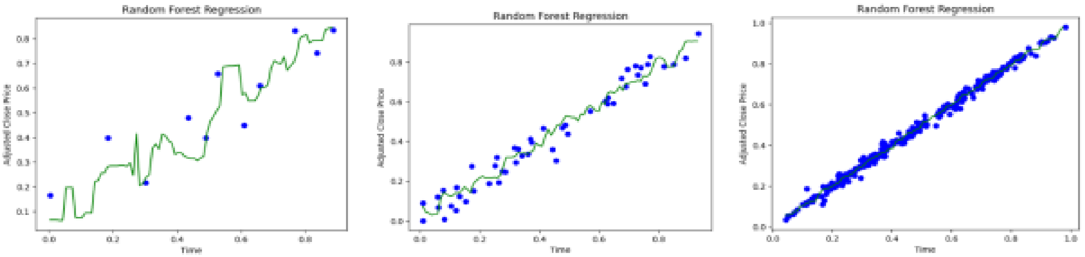


Figure 5: Regression plots of Weekly, Daily, and Hourly data (from right to left)

The last model we implemented is the Linear Regression – Normal Equation method. The Normal Equation method is a one-step algorithm used to analytically find the minimizing coefficients of the loss function. The preliminary model's performance will be used to predict the adjusted closing prices of the last 10 weeks, days, and hours of our datasets. The metrics of success used for this model will be the MSE and the R^2 values. Weekly data resulted in an MSE of 0.0118 and an R^2 of 0.7463, Daily data resulted in an MSE of 0.0027 and an R^2 of 0.9458, and Hourly data resulted in an MSE of 0.0003 and an R^2 of 0.9923. These results were consistent with the rest of the models but were the best results out of all four of the models.

Finally, after testing all four models, the Linear Regression – Normal Equation method performed the best for all three datasets. Implementing this model into the 'real world' means that we will be using data outside of the time frame we selected for our dataset, which we can check in real time to see the accuracy of the model's performance. The best data to use for the features selected for this project would be the daily data, since it had the best success metrics and had the lowest likelihood of overfitting. The prediction vector yielded from the model contains the predicted stock prices for each time period in the dataset based on the input features used, i.e., the volume of shares and the open, high, and low, prices of the stock.

One challenge we faced during this project was feature engineering. We were interested in looking into moving averages, relative strength indicators, and momentum indicators to get more insight into a certain stock performance relative to its competitors and the direction a stock is likely to move in based off its performance in the past. We had trouble implementing this into our final analysis in the given time constraints, so this would be something we would like to analyze in the future.

In conclusion, we noticed a significant trend with the models' performances. The weekly data performed the worst, while daily and hourly data had much better performance metrics. Given the scope and features of our data, stock price and volume are more suited to predicting closing stock price on a daily and hourly basis. We believe this is due to the volatility of the market and how quickly things can change in a day, let alone a week, within the stock market. In our real-world test of the model, since these closing prices have not yet happened in the market, we will only be able to tell the accuracy of our predictions in the days to come.