Niam Bashambu, Adam Ancheta, and TJ Kalapatapu

Elena Strange

DS 2500

March 1, 2024

What is investing? What stocks should I buy?

If considerations in a dataset of stocks such as lowest value, lowest average value, highest value, percentage growth/decline per year, and k-means clustering are calculated from this project, this will allow for investors to buy stocks and make a return based on the history of these stocks. Based on this data, we will be using different ML models to predict future returns. Additionally, an important attribute to the ML model is using regressions being created to predict the next day closing price of the stock.

As people who are constantly searching for ways of creating wealth for ourselves and future generations of our family, we found that the stock market and investing was extremely appealing. Many investors who are new to the stock market have a burning curiosity and concern about which stocks to invest in. From recent trends and many different sources, there will be recommended stocks to buy and not one to make a decision based on historical data. Due to the volatility of the stock market, this project attempts to use the amount of investors and how this has an influence on the individual stocks, while also calculating the impact of the entire market from past data. Using past data and statistical analysis not always, but sometimes able to show us the possibilities of the future. Not only do we bring the relevance of the project into this section, but this outlines the entirety of the project. Something that we could work on is stock analysis and creating a project that takes the input of a certain stock or index and based on analysis levels of lowest price or highest price or some other feature of statistical analysis over the years in order to see if the stock is a worthy investment. We are planning to do a project about the Stock Market from 2014 to the present day (Feb 2024). Throughout the entirety of this project, this helps to analyze stocks and build generational wealth or financial freedom for individual or future generations.

Yahoo Finance, a reputable and trusted source that tracks stocks, news, and information pertaining to the ambiguous realm of finance. As one of the most prominent sources, this has provided information and historical data of the stocks in this project. The data set includes the highest closing price and our group has edited this to be able to analyze every stock that is given. The data set is from 2019- 2024 of 10 stocks (Uber, Adidas, AMD, Intel, Apple, Google, Meta, Nvidia, Microsoft and Nike), which highlights the realistic trend and growth, as sometimes going back to when the company was small is not as relevant. Balancing enough historical data and recent data allows for a more realistic prediction or analysis of the stock, therefore a better recommendation will be given from the ML model we created. The data was collected by downloading the file from the website and creating a csv, then we worked from there. Some of the ethical considerations that we came up with were the privacy of customers data, privacy of information, potential insider trading or illegal tips affecting price, data being manipulated on this website, data being faked, and many other factors that are up for questionable ethics. From this idea of ethical actions, the data can definitely be biased or target a certain audience. Many of these large corporations have this idea of capitalism and do anything for more money. Due to this mentality of profit and money greedy companies, the data could be biased for those who have money and are interested in the idea of investing in corporations.

One of the most important attributes and data science approaches to this project were the python libraries. There were about 9 python libraries and there were about 7-8 extensions of one python library. Some of the most important python libraries included the csv library, Pandas library, global library, and the Sklearn library. The csv library was used to convert the Yahoo finance data into an acceptable file type that is able to be edited by python. Additionally, it creates a place for storage of each individual stock data that is partially organized. After using the csv library, we moved onto the Pandas library, which is the library that dropped the irrelevant stock data and allowed for a cleaned csv file of each individual stock. Additionally, some of the plots emerged from the Pandas python library, such as the linear regression and logistic regressions of each individual stock. Onto the global library, this helped the project by creating a file path to the repository of csv files, which allows us to do the analysis and regression of each stock. Finally, the Sklearn library is an important python library, as the predictive analysis and regression plots for each stock were created. It created different training and test data that recognizes a pattern for the data and makes a predictive assumption of the closing price of the stock the next day.

The probability prediction is used to map the future outcomes of the stock that is selected from the data set. The ML model or python Sklearn library has a built-in function that predicts the probability based on the average of the data set. Pertaining to this project this finds the average probability of the stocks percent growth or decline for the day and uses that relationship to map the future probability. By adding a linear fit to the probability prediction value, this allows for a constant probability to map the likelihood of that event happening. Additionally a feature in this code was created, which was the average predicted change and is compared to an investment threshold. By creating an investment threshold this will allow for a decision of investment to be made for either a promising or downward trajectory.

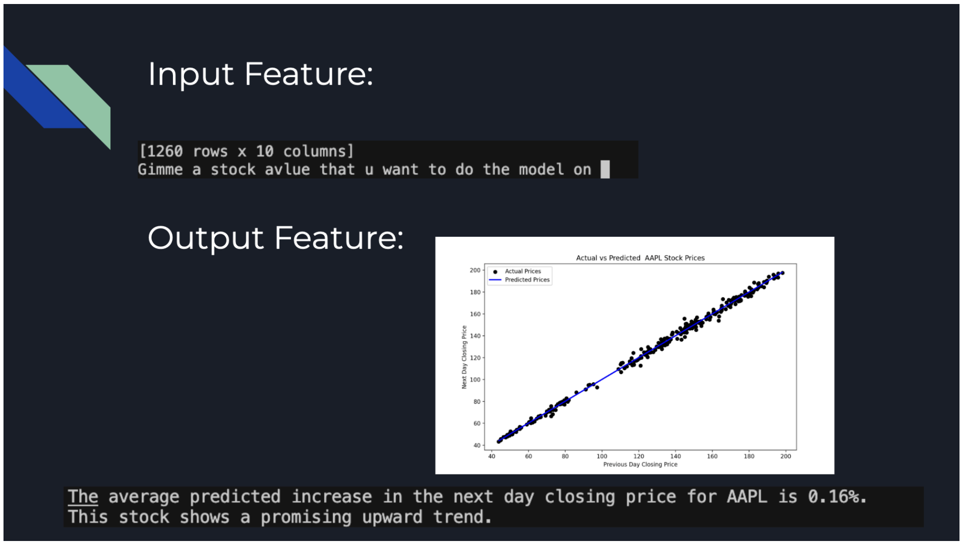
The logistic regression in this project is used to map the relationship between the predicted values and the previous day closing price of each individual stock. Finding this relationship will allow for the trends, growth rate, decline rate, and a predictive probability of certain events happening to be seen. By doing this many analysts and businesses can use this predictive model to map the probability of price increase or decreasing, prompting a decision to be made by a human and not the computer.

The linear regression is a classic regression, as this type of plot allows us to see the constant rate or average growth of a stock assuming it almost always increases. The line of best fit for the linear regression in this project maps the actual closing price per day versus the next day closing price. By mapping these two variables we can see the rate at which a stock's price actually increases or actually decreases based on the probability prediction with the ML model.

K means clustering is used to create a generalized category or cluster of many data points that are similar to each other. For this project, we utilize the elbow method and this allows us to find the optimal k value or optimal probability and actual closing price, which will allow us to make smart and informative decisions. K means clustering seems important because if the actual outcomes and predicted outcomes are mapped in a similar place. This means that the k value is optimal and the percentage of accuracy is far greater for mapping future possibilities.

All of the different figures demonstrate the output and analysis that this project has done. Whether it was taking a user input and returning values or an opinion to see if the stock is increasing and using predictive regression models in order to prove the purchase of the stock.

Figure 1: Input/Output Feature:



Input Feature:

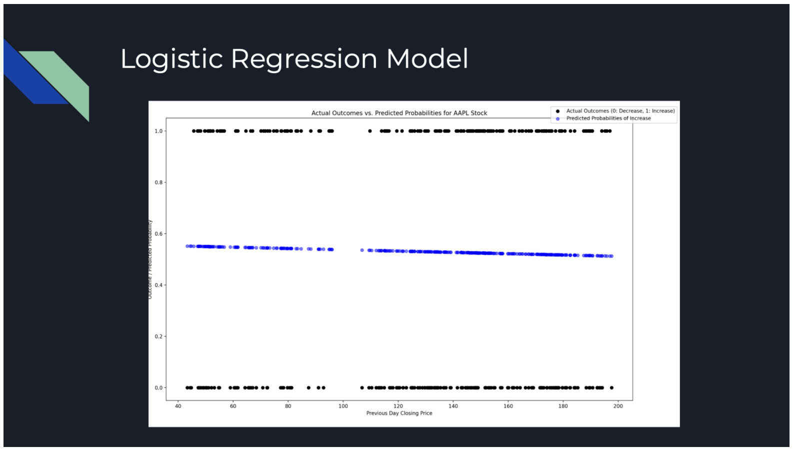
```
[1260 rows x 10 columns]
Gimme a stock avlue that u want to do the model on
```

Output Feature:

```
The average predicted increase in the next day closing price for AAPL is 0.16%.
This stock shows a promising upward trend.
```

Figure 1 is an important part of this project as it prompts the user into selecting a stock within the database. After doing this it outputs the linear regression graph which illustrates the predictive model of next day closing price and previous day closing price. This is important as it maps a generalization of potential growth rate for stocks. Not only does this return a linear graph, but this also returns the average predicted increase for the next day of apple which is 0.16% total increase and the decision if the stock is upward trending. This is one of the most prominent parts of our project, as it makes a decision based on any data that is given in a cleaned csv, which will allow users or analysts to make decisions on stocks.

Figure 2: Logistic Regression model:



Logistic Regression Model

Figure 2 is an important model for another aspect in this project, as it demonstrates the relationship between the expected values and the closing price of each individual stock the day before is mapped out using logistic regression. It demonstrates multiple factors such as the trends, growth rate, decline rate, and a prediction chance of specific occurrences occurring can all be observed after this link is established. For this graph of the AAPL stock, or in other words, apple is a company that is clearly on an upward trajectory, however, for the period of time this graph demonstrates that the probability of apple closing price to be higher gets slightly lowered from about .58 to .5. By using this regression analysis we are able to see that Apple has a 50% probability of either going up or down for the next day.

Figure 3: Linear Regression using Predicted Values/Previous Day Closing Price:

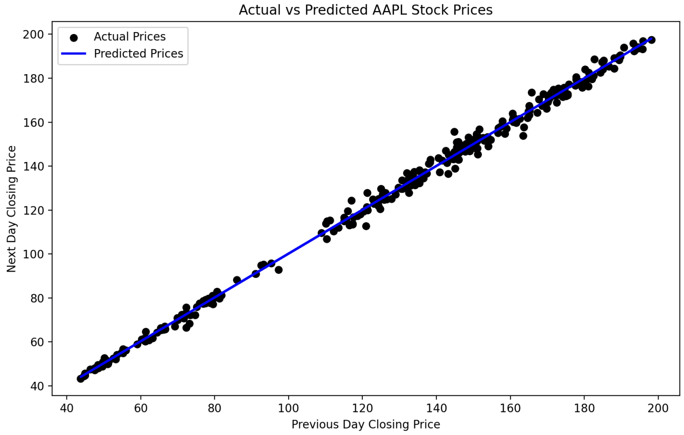


Figure 3 demonstrates linear regression is a traditional regression since it shows us the average growth rate or constant rate of AAPL and every other stock in the data set.. The real closing price for each day compared to the closing price the following day is plotted on the linear regression's line of best fit for every stock.

Figure 4: Elbow Method K Means Clustering:

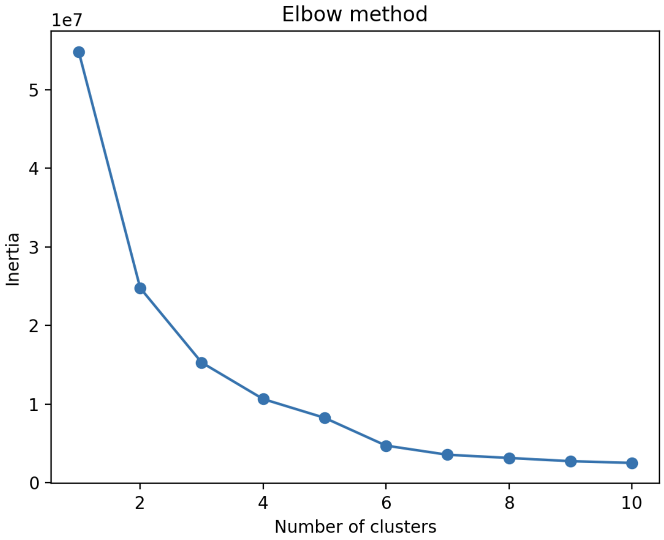


Figure 4 demonstrates that as we increase the number of clusters the inertia or how far the k value is away from each cluster goes down or becomes closer. Here this is the elbow method upon the AAPL stock which demonstrates that as we raise the number of clusters or more data points, that the amount distance between a centroid and a data point nearby will be lower. This will allow investors to generalize prices and assume that at that certain price there will be either an upward trend or downward trend.

Throughout the entirety of this course, we have learned alot about the attributes that are essential to the field of Data Science. DS2500 has been an important course that covers a lot about coding ML models in python and utilizing certain libraries that are useful in processing and analyzing data. Not only did we learn alot about ML models, but we learned how to apply classroom lectures into real world projects that will one day make an impact on the world. One of the most significant aspects of learning throughout the entirety of this project was the ability to code and the power of python libraries. Learning how to do this right now will not only help us for this class, but in our potential careers as Data Scientists. The significance of this project spreads widely, as many people are new to investing and wondering if there is a model that can precisely pick upward trending investments. Oftentimes many people want to build generational wealth for their families and gain financial freedom and invest in companies that will grow wealth over a long period of time.

Some of the future that would be included for this project is to explore a diversity of stocks and a diversity of different investments. We could explore an ETF, Index, Bonds, or some other form of investments. We believe that there should be more of an emphasis upon diversity and analyzing the return of different investments would allow for a better understanding of what is a worthy investment. Not only would we want to include other types of investments, but also potentially something like an informational source or database could be created as an extension to understand everything you need to know about investments. By doing this, it will secure that you and your family will build wealth that could be used through multiple generations.