**Milestone 3: White Paper Report**

Jamie Tran

Bellevue University

DSC680 Applied Data Science

Professor Amirfarrokh Iranitalab

**Business Problem:**

The business problem we will be tackling in this final project will be focusing on generating more income by retaining/attracting new clientele for investment firms such as Fidelity Investments or Charles Schwab by providing resources to their clients to help them be more successful. In businesses such as these they want clients to open up accounts with them to house their investments. New and old investors are always looking for strategies to help the grow their accounts. Trustworthy analysis sentiments that give insight into bullish and bearish stocks will help us reach new clients that may be novice to investing as a whole as well as, supply our existing clients with an outpour of updated insights.

**Background/History:**

When you enter adulthood people are always telling you to be smarter with your money, investment your money and hopefully one day you will be able to retire early. In reality most people truly do not understand the stock market or the positions they hold in their accounts. According to nerdwallet.com only 43% of people with workplace retirement accounts know all of the investments they own in their accounts. That number drops to 38% when it comes to those who own nonretirement investment accounts. From these statistics it is clear that most people lack the knowledge to properly manage their investment accounts on their own. Our model will be a large contributing factor to help us guide new and ‘experienced’ investors. Ridge regression analysis will be able to predict the future outlook of stocks to include in a sentiment analysis report. Considering that 44% of Americans choose stocks to invest in based on their future potential growth this tool will help us stand out from competitors.

**Data Explanation (Data Prep/Data Dictionary/etc.):**

The datasets that I have chosen were selected from yahoo finance. Each dataset contains 3 years’ worth of historical data for each stock. The data frame is made up of 7 columns that include the date, opening/closing prices, high/low values for the day, adjusted closing price, and volume of stock sold. Since we are gathering this historical data from a reliable source the data preparation will be more condensed than other projects we have completed. To begin the preparation, I began by checking each data frame for any missing values. Fortunately for us there were no missing values so no additional steps were required. From there we only needed to remove the date column from all datasets so we would be left stock specific data to perform analysis on. Once the column was dropped from the data set for each stock the data frames were split in two. A x dataset was created with all columns apart from the ‘close’ column and a y data frame was created from the values in the closing price column. The final step in the data preparation was splitting these new data frames into training and test sets for each stock.

**Methods:**

I will begin this research with exploratory data analysis to discover any underlying trends and make a few initial call outs. Once that is completed and my data has been cleaned and prepped, I will move into building a ridge regression model. Ridge regression is an extension of linear regression which will be ideal for our research because there are a handful of variables from our dataset that are essential for predicting future closing prices for each stock.

**Analysis:**

To kick off the exploratory data analysis I created the two visualizations included below. The first graph displays a histogram showing us the volume of closing prices for all three stocks over the last three years.

A graph of a stock price

Description automatically generated with medium confidence

This graph shows us that Google has the lowest stock prices while Tesla stock has proven to be the most expensive over the years, ending with Apple smack dab in the middle.

A graph of a graph of stocks

Description automatically generated with medium confidence

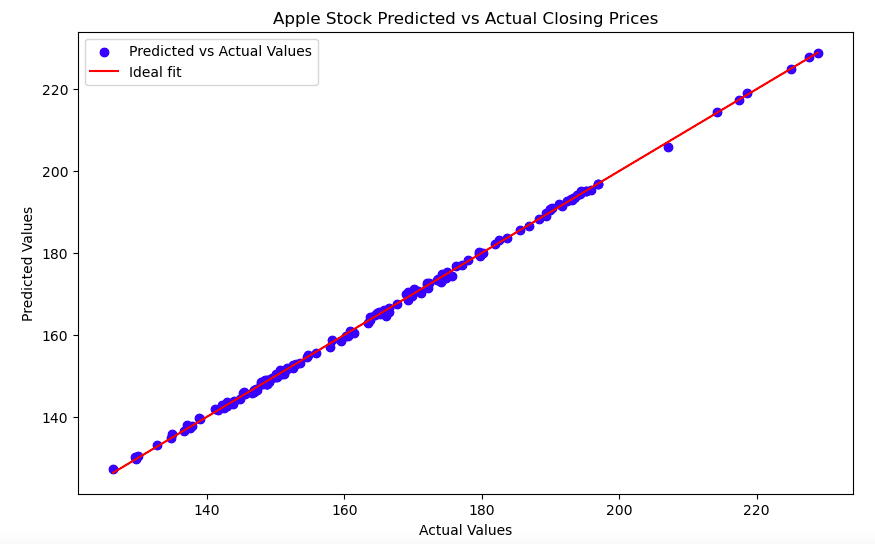
This next histogram shows us the volume of stock sold each day over the past three years. You might assume that since google has the cheapest stock that they would be selling more units of their stock. However, this graph surprisingly shows us that although Tesla has the most expensive price out of the three and is still the most sold. Similar findings with Google, we can see that Google has sold the least number of shares per year all while being the least expensive. This would indicate to us that the stock may be the least sought after out of the three.

After the EDA was completed, the data was checked for missing values. From there we had split each data frame in two. The x data set contained all of our independent variables such as the opening and adjusted closing price, high and low prices for each day, and the volume of stock sold each day. Our y data set however just contained the closing column for each stock. The final step before performing our analysis is splitting our x and y data sets one step further into training and test sets. After we have established our training and test sets, we will fit and transform the x train and test set using the standard scaler. Finally, ridge regression was ran using the y train set and the scaled x train set.

To evaluate our model, we will be creating predictions for the scaled trained and test sets. Once those values have been generated, we will calculate the mean squared error and r-squared value using the training and test predictions with their corresponding y training and test set. The r-squared values for all three stock produced results of 0.99 for both the training and test sets. Considering that the strongest value we can get for this metric is one we have been able to produce very strong results for all three models. This lets us know that our independent variables were able to explain 99% of the variability in our closing value. Our model was also able to produce a strong mean squared value for all three stocks. Google had a mean squared error of 0.12-0.13, Apple’s MSE ranged from 0.29-0.31, and Tesla’s values were between 0.63-0.66. The mean squared error value lets us know the average squared difference between are predicted values and the actual values. Which in our model would be anywhere from $0.12-$0.66, considering our stock prices are upwards of $100 the error amount is miniscule.

Now that we know our model is reliable and has produced accurate values, I have created the following graphs to display whether or not the outlook for the stock is bullish or bearish. These graphs as well as the outlook are valuable information that will be included on our sentiment analysis we are trying to build.







From the following graphs we can clearly see that the outlook for Google, Apple, and Tesla stocks are bullish.

**Conclusion:**

Based on the results for our model I believe our current model is ready to be utilized when building a sentiment analysis for investment firms. Our evaluation metrics have proved that our model is reliable while maintaining high accuracy. With this model we will be able to provide great insights to help new and old investors looking for insight on their next move. However, this model should not be utilized on stocks with high volatility such as penny stocks. While well-established investment firms should know better than to provide definitive outlooks on stocks such as these there might be smaller firms who try to adopt this model as a one size fits all. Due to the unpredictable nature of these stocks, they are not suitable for this type of model.

**Assumptions:**

An assumption that we are making with this model and with all linear regression models are that the values provided do not change. We are assuming that historical data provided will be similar or the same variables in the future. This assumption that the observations are separate from one another is a clear difference between our model and a time series model. The last assumption we are making is that the relationship between the closing value and the independent variables are linear. While this might not be ideal to predict the swings of stock prices it is prefect for showing us the outlook of a stock.

**Limitations:**

A limitation of the linear regression model is it doesn’t take in how these values might change over time. This model assumes that the relationship between the independent variables and the closing price can be explained with a straight line. Each value in the dataset provided is treated as an independent value, because of this the natural dips and swings of the market are not accounted for when performing this analysis.

**Challenges:**

As mentioned previously, this model would not be ideal and will likely produce less than desirable results for highly volatile investments such penny stocks or crypto currencies. Another challenge of this research is there are always unknown variables at play, variables that may not even exist at the time of research. This can create issues because uneducated investors can take results that differ from our analysis as us scamming or setting them up for failure. This model only uses the history of the values to produce future values. Therefore, other very important independent variables such as the stability of government is not factored into this research.

**Future Uses/Additional Assumptions:**

In the future this model will be a great deal of help when helping data teams determine the outlook of stock performance. This would be one of the main models being used to build a sentiment analysis that can be made accessible to investors. The analysis report would contain baseline information such as the bullish or bearish outlook on specific stock. I see the stocks used for this model likely being shares well known to the public that investors want to know more about. However, the reality of the stocks chosen will ultimately be the ones deemed most important by the firm.

**Recommendations:**

Before making this model available to companies I would recommend testing it on a few other smaller stocks. The positions selected for this research were picked because new investors will likely be aware of these big three shares and wonder which one would be best and have the ability to compare them. However, there are stocks with less historical data or far lower prices than the three researched for this project. Ensuring that accurate results can still be produced for mid-size stocks is just as important.

**Implementation plan:**

I see this model being implemented by investment firms such as Fidelity Investments or Charles Schwab. These large investment companies have a user interface that is widely accessible to the public as well as tools that are only available to those who have accounts with them. The firm can choose to make only a portion of the report available to the public to draw in new investors. Once they have established an account with the company, they client can have access to the full report.

**Ethical Assessment:**

There are many ethical considerations to be aware of when dealing with people’s money. The most concerning scenario that we need to consider is the possibility that our analysis turns out to be vastly different from reality, poor results like this could cost our clients hundreds or thousands of dollars. With this being the case, we need to ensure integrity on our end by verifying that our model is producing strong reliable results before handing this information over to our clients. We must also encourage our clients to do their own research and give them the proper precaution/knowledge to inform them that even with the best analysis, the stock market is never guaranteed and there will always be a risk that they can lose on their investment. The goal of providing them with this discretion is to prevent them from jumping blindly into purchasing a various stock or putting in more money than they can afford to lose.

**Ten Questions:**

1. Does this sentiment analysis guarantee results to investors?
   * Specific results can never be guaranteed when it comes to investments, there will always be variables out of our control. The sentiment analysis is intended to provide some direction for investors looking for more guidance.
2. How can we trust the results of this research?
   * Our evaluation metrics produced results that were able to provide us with the reassurance that our model is reliable and accurate. Based on our results the independent variables chosen explain 99% of the variability with the closing price. The average mean squared error also lets us know that the average variation between our predicted values and the actual values for the Google, Apple, and Tesla holdings are only $0.12-$0.66.
3. No one can predict the exact values of the market so why are you claiming to?
   * We are not stating that we can provide exact values with our model. It is virtually impossible to predict the exact prices of stock. However, we are providing the most accurate results we can produce on the outlook of these positions based on the past performance.
4. Will this tool only give us the results for three stocks at a time?
   * No, this model can be performed on a variety of different investments.
5. Will we only receive the outlook of the same three stocks?
   * That is not the intention of this model, ideally a plethora of positions would be run using this model to provide a diversified sentiment analysis report.
6. Why would investors use this when they can just google different articles?
   * The advantage of this report is to have a centralized resource of trustworthy data. Going back and forth between a number of websites and Google searches that may not even contain accurate information can be tiresome and confusing.
7. How often can this report be ran?
   * There is no limit to how often this report can be run. It is up to the investment firms and the cadence they want to uphold with this information.
8. How long does this model take to produce results?
   * The time taken to run this analysis can vary with the size of data provided. It can be anywhere from a few seconds to a few minutes.
9. Is there a limit on how much historical data can used to generate results?
   * There is no limit on the amount of historical data that can be provided. Two things to consider with the data size are the results are likely be more accurate with more data, however it could take longer to run.
10. Will this report provide closing values for each day of the upcoming year?
    * This model will produce real values however, the intention behind running this model is to look at the overall direction of growth as whole rather than the individual values.

**References:**

“Do You Know What Your Investments Are? Most Americans Don’t.” *NerdWallet*, www.nerdwallet.com/article/investing/do-you-know-what-your-investments-are-most-americans-dont#:~:text=Key%20findings,they%20hold%20in%20these%20accounts. Accessed 31 July 2024.

*Alphabet Inc. (GOOG) Stock Historical Prices & Data - Yahoo Finance*, finance.yahoo.com/quote/GOOG/history/. Accessed 31 July 2024.

*Apple Inc. (AAPL) Stock Historical Prices & Data - Yahoo Finance*, finance.yahoo.com/quote/AAPL/history/. Accessed 31 July 2024.

*Tesla, Inc. (TSLA) Stock Historical Prices & Data - Yahoo Finance*, finance.yahoo.com/quote/TSLA/history/. Accessed 31 July 2024.

**Appendix:**

**First ten rows of the google dataset:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Adj Close | Volume |
| 7/30/21 | 135.511002 | 135.771347 | 134.814194 | 135.220993 | 135.067291 | 23954000 |
| 8/2/21 | 135.484497 | 136.020493 | 134.669495 | 135.989502 | 135.83493 | 20140000 |
| 8/3/21 | 136 | 136.335495 | 134.183502 | 136.279999 | 136.125092 | 19064000 |
| 8/4/21 | 136.249496 | 136.537994 | 135.414993 | 136.028503 | 135.873886 | 16528000 |
| 8/5/21 | 136.028503 | 136.949997 | 135.600006 | 136.940002 | 136.784348 | 11866000 |
| 8/6/21 | 136.294998 | 137.080048 | 136.046494 | 137.035995 | 136.880234 | 13560000 |
| 8/9/21 | 136.949005 | 138.321503 | 136.428757 | 138.001999 | 137.845139 | 12374000 |
| 8/10/21 | 138.079498 | 138.551498 | 137.226501 | 138.096497 | 137.939529 | 16034000 |
| 8/11/21 | 138.283005 | 138.847748 | 137.350006 | 137.689499 | 137.53299 | 15204000 |

**First ten rows of the Apple dataset:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Adj Close | Volume |
| 7/30/21 | 144.380005 | 146.330002 | 144.110001 | 145.860001 | 143.408936 | 70440600 |
| 8/2/21 | 146.360001 | 146.949997 | 145.25 | 145.520004 | 143.074615 | 62880000 |
| 8/3/21 | 145.809998 | 148.039993 | 145.179993 | 147.360001 | 144.883698 | 64786600 |
| 8/4/21 | 147.270004 | 147.789993 | 146.279999 | 146.949997 | 144.480576 | 56368300 |
| 8/5/21 | 146.979996 | 147.839996 | 146.169998 | 147.059998 | 144.58876 | 46397700 |
| 8/6/21 | 146.350006 | 147.110001 | 145.630005 | 146.139999 | 143.89949 | 54126800 |
| 8/9/21 | 146.199997 | 146.699997 | 145.520004 | 146.089996 | 143.850235 | 48908700 |
| 8/10/21 | 146.440002 | 147.710007 | 145.300003 | 145.600006 | 143.367752 | 69023100 |
| 8/11/21 | 146.050003 | 146.720001 | 145.529999 | 145.860001 | 143.623779 | 48493500 |

**First ten rows of the Tesla dataset:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Adj Close | Volume |
| 7/30/21 | 223.919998 | 232.509995 | 223 | 229.066666 | 229.066666 | 88969200 |
| 8/2/21 | 233.333328 | 242.313339 | 232.800003 | 236.556671 | 236.556671 | 100847400 |
| 8/3/21 | 239.666672 | 240.883331 | 233.669998 | 236.580002 | 236.580002 | 64860900 |
| 8/4/21 | 237 | 241.633331 | 236.309998 | 236.973328 | 236.973328 | 51007800 |
| 8/5/21 | 238.666672 | 240.316666 | 237.136673 | 238.210007 | 238.210007 | 38758800 |
| 8/6/21 | 237.300003 | 238.776672 | 232.543335 | 233.03334 | 233.03334 | 46869000 |
| 8/9/21 | 236.723328 | 239.676666 | 235.043335 | 237.919998 | 237.919998 | 44145900 |
| 8/10/21 | 237.996674 | 238.863327 | 233.960007 | 236.66333 | 236.66333 | 40296900 |
| 8/11/21 | 237.570007 | 238.393326 | 234.736664 | 235.940002 | 235.940002 | 29401800 |