Christian Peterson and Ivan Sapozhnikov
DS2500 Fall Semester Final Project
5 December 2023

Stock Price Prediction & the ARIMA Model - An Analysis

Problem Statement and Background

Predicting stock prices is an extremely daunting task that, per common opinion, is near-impossible. However, it is in high demand because of the high returns it potentially can offer. The reasoning behind this is that stock investments, in comparison with other investments - bonds, CDs, and such - offer the highest average returns. For example, investing in the S&P 500 index has provided investors with an annual return of 10% on average (Royal, 2023). The risk-return relationship, in which investors can theoretically acquire higher returns or losses when choosing investments that are considered riskier, serves as evidence of the necessity for more accurate stock price predictions, as the greater the likelihood of the measurement, the more likely that a person will be able to have an idea of the return that they may receive.

However, there are limitations and caveats to this statement; greed and irrational behavior have the potential to undermine the performance of any stock prediction method or model. This can be represented by the nature of behavioral finance theory, which is based around how irrational behavior, systemic patterns of behavior, and limits to arbitrage in financial markets hinder the ability of the market to reflect economic fundamentals (Goedhart, Koller, & Wessels, 2005). Without getting too technical, systematic patterns of behavior, for example, connect to the theory that "patterns of overconfidence, overreaction, and overrepresentation are common to many investors and that such groups can be large enough to prevent a company's share price from reflecting underlying economic fundamentals," (Goedhart, Koller, & Wessels, 2005). This statement is interesting in accordance with the theory in the prior sentence: both hint at the fact that current share prices are not reflective of economic conditions and determinants, and an inference of this factor is that share/stock prices are difficult to hence predict in the future. The reasoning behind this is that, if a current value or measure is not fully backed by fact or science, then it is also backed by "noise." This is an issue because, in any prediction model, the effects of "noise" essentially need to be smoothed out in order for an accurate future value to be predicted, as this allows for specific patterns to be clearer (not obscured by externalities). These factors are interesting in the case of the stock market in their ability to contrast and complement the Random Walk Theory of Stock Prices, which "casts serious doubt on many other methods for describing and predicting stock price behavior-methods that have considerable popularity outside the academic world," (Fama, 1965).

With these cases hindering the ability of the market to reflect economic fundamental ideals and conditions, the potentiality of accurately predicting stock returns becomes extremely difficult and, in the case of the Random Walk Theory, is disputed and speculatively disproven. The first objective of our report, therefore, is to gain a better understanding of the extent of these theories. We want to measure the effects to which a share price, or return, model can predict the next occurring price in the future; we want to analyze our model's performance, during training, on how accurate its predictions were with respect to actually-recorded prices, and then eventually predict the share price for the n+1th day. We hypothesize that the model will produce more accurate predictions for stocks in specific industries, especially those that are less volatile than others, but this is to be discussed further.

There is also another issue that we believe our project should address and explore further. Predicting stock prices also requires powerful and pristine technology to handle copious amounts of data. Especially in terms of automating the prediction process, obtaining the highest accuracy is based on having the highest data volume and quality, among other factors such as mitigating non-linear external influences, avoiding overfitting the model, and preventing the lack of transparency with the obtained results (Avci, 2023). For this reason, we hypothesize that our predictions, when testing the model, will not be close to being fully accurate not only because of the difficulty of doing so but also because of the quality of our obtained data. We predict that our model will be somewhat accurate for short-term data and will not be effective for the prediction of any share price after the n+1th day.

Introduction to the Data

Our data source contains spreadsheets that utilize stock share price and other measures, obtained from Yahoo Finance through the unofficially-connected `yfinance` API. There are five sets of spreadsheets that we use, all belonging to one specific industry (all chosen at random); these include the Technology, Automotive, Health Sciences, Oil, and Filmmaking sectors of the global economy. However, some of the companies in each industry exert a larger influence on the American economy in comparison to other economies, or the global economy, but this is irrelevant in terms of our project and its goals.

In each industry, ten specific companies of which are publicly traded in global financial markets or OTC (over-the-counter), have been tested. The data from the test for each company makes note of its actual daily closing share price, its predicted daily closing share price, and yesterday's daily closing share price (in this order in each spreadsheet). However, there are missing values, due to the fact that the markets are not open during weekends, so there is no recorded closing share price. However, our model still predicts values for these days to account for movement occurring due to effects from the media, news, and economic effects. Once all companies of one industry have been modeled, one final spreadsheet containing the expected return determined by the model, expected return without daily trading, mean return from randomly buying and selling the share(s), the standard deviation of the random return from the random buying and selling, and the maximum return gained from the random buying and selling all over a specified time period of one year is added (to that specific industry's file). This process is done for each of the five industries that the project has been designed for.

There are no privacy or ethical issues to be concerned about with our data and the collection methods used to obtain them. The reasoning behind this is that the root of where our data stem from, `yfinance`, is an unsponsored library providing data from Yahoo Finance. The data associated with Yahoo Finance does not have the potential to uncover privacy issues, or illegal/legally unobtainable data issues, and for that reason, there should not be any concerns associated with the data.

Bias does exist in our data. However, it must exist due to the nature of stock share prices and trading. For example, many investors trade shares based on emotions, short-term developments and changes, and irrational thinking that “can lead to poor decision-making and adverse portfolio outcomes,” (Morgan Stanley, 2023). In terms of bias due to missing information, there is no issue, as the retrieved data has small amounts of missing data but not an overwhelming amount. If this was not the case, there would be a severe issue as the lack of relevant information would therefore skew our dataset and produce extremely inaccurate predictions.

Data Science Approaches

The data science approach apparent in the project is based on using the Autoregressive Integrated Moving Average (ARIMA) algorithm to train, test, and assess a model's performance in predicting the stock share price for the n+1th day, among other tasks. The ARIMA statistical analysis model is the only main algorithm that we used to obtain our project findings. For this reason, it is necessary to explain how the process works and can be applied to the obtained data.

The ARIMA is based on three calculations, or pieces, in order to set up a valid prediction model: the Autoregression (AR), Moving Average (MA), and Integration (I) portions. The first part, the Autoregression, represents regression with past values within the same series, where each piece is correlated with every piece in the past and their errors; in the model, this part is represented by a “p” which serves as the number of periods in which the prediction is dependent on. In other words, the Autoregression is weighed by the correlation between past values. The second part, the Moving Average, represents the fact that the present value in a time series model is dependent on its past values and the error associated with those. In a larger significance, the Moving Average holds the weight of the correlation with the past errors, and the term “q” represents the order of that model. Finally, the Integration part of ARIMA represents the differencing that occurs to remove the trend and seasonality from the model, with a “d” representing the number of times the data is being differenced; the differencing allows the time series data to become stationary, which is important because this is a primary assumption of the ARIMA model. The seasonality and trend patterns are not being removed from the data, but are rather being separated from the random errors.

Before the model can be constructed, an ARIMA "object" needs to be defined, based on a specific order of "p,d,q". The determination for which order of "p,d,q" that needs to be chosen is based on the analyses of two plots, the Autocorrelation Function Plot (ACF) and the Partial Autocorrelation Function Plot (PACF). A feature of these models is the Akaike Information Criterion statistic (AIC), which represents a combination of the model's simplicity and parsimony, and goodness of fit. When determining the different ARIMA orders, the lower the AIC is, the "better" the model will be in making a prediction. In general, the best values of "p,d,q" should ideally correspond to the lowest value of the Mean Squared Error, but this is not always the case when choosing the smallest AIC value. It is always wise to choose the model with the lowest AIC, but it is important to note that the smallest AIC does not always correspond to the model with the smallest Mean Squared Error. For this reason, it is important to implement two methods of assessing the models to influence the likelihood of obtaining a more accurate prediction: automation and self-checking (self-checking both the automated response and all of the AIC statistics).

The ARIMA model needs to be constructed on data designated as the training set, which makes up the larger portion of the obtained data and tested using the data designated as the testing set, which makes up the smaller portion of the obtained data. The testing leads to an analysis, or model evaluation, which is completed by comparing the model's prediction of values corresponding to dates in the testing set to the actual values recorded on those dates. After the model has been evaluated and has been deemed effective, it can then be used to predict future values. These values can, of course, extend to longer future forecasts, but there is a high risk of the prediction model becoming highly inaccurate (Bora, 2021).

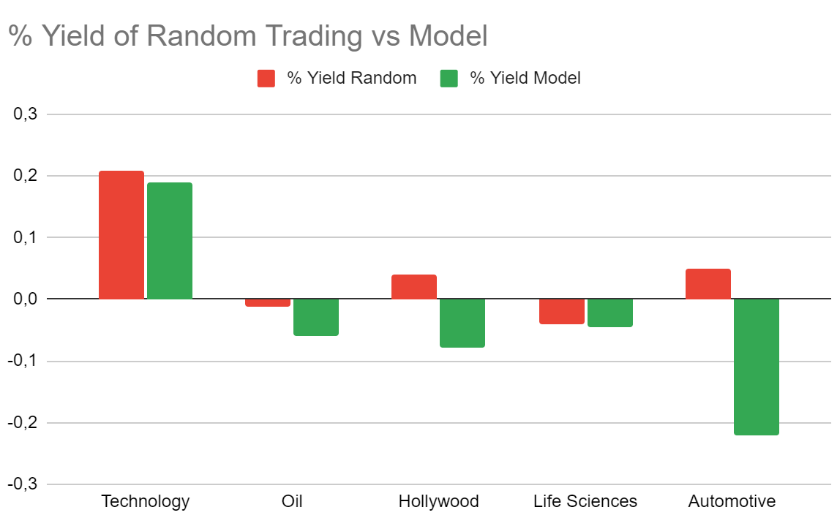
After all of these requirements have been considered, checked, and implemented, the model can be used to make predictions and serve a purpose.

Results and Conclusions

Our results stem from two analyses. Firstly, the model's prediction performance in terms of predicting the testing set values, comparing them to yesterday's stock share price and measuring them in relation to the actual occurrences. Secondly, the model was applied using a robo-trading strategy where, if the predicted share price for the day was greater than yesterday's price, the model would purchase one share and sell it on the following day; if the opposite had occurred, the model would hold the stock (strategy: longer than day trading, shorter than swing trading). After the predetermined period has passed, a returned value would specify the total amount of money lost or gained using the automated trading strategy. This value is compared to the expected return without daily trading (holding) and the maximum, mean, and standard deviation return from randomly buying and selling.

This is all being done to test our hypothesis that the model will predict more accurately for specific industries, in comparison to others. This hypothesis is based on the nature of different industries and the volatility levels of their returns (Speights, 2023).

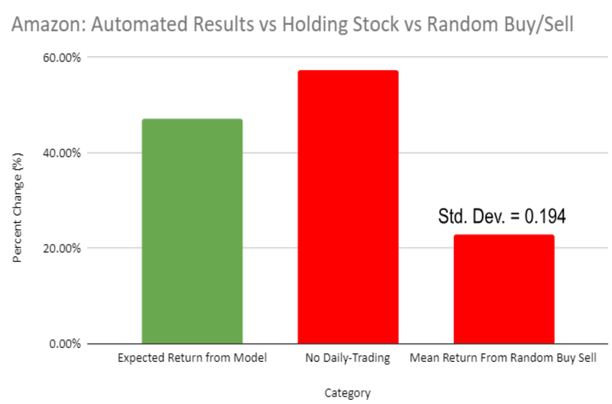
<u>Industry-Wide Visualization:</u>

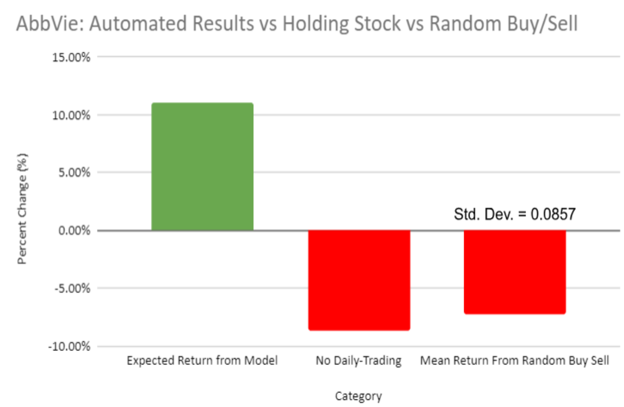


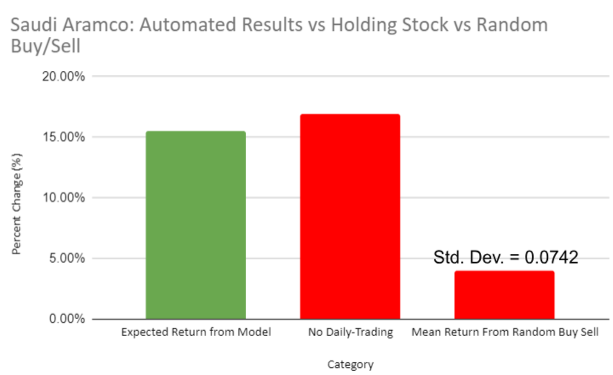
<u>Analysis:</u>

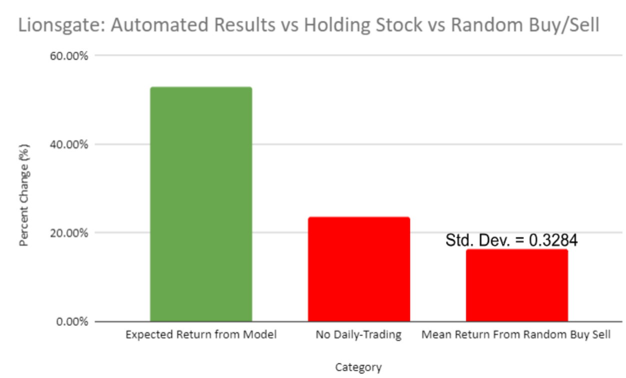
This visualization conveys the model's accuracy on a per-industry level. There are two metrics implemented: the percent yield/return of randomly trading without relying on any model predictions (red bar), and the percent yield/return from the model's robo-trading function which relies on its predictions (green bar). When analyzing the difference between the two percentages, across all industries, it can be seen that robo-trading mechanism incorporated in the model is always being outperformed by that of randomly trading; this proves that our model is therefore unable to capture any effect of one industry (or more) being more accurately predicted in comparison to another (others). In short, the model has performed, and predicted, poorly.

Specific Companies Visualization:









Analysis:

This visualization analyzes the model's accuracy on a per-stock basis and hones in on the outliers that were captured in the dataset: Amazon, AbbVie, Saudi Aramco, and Lionsgate. These companies' stocks hail from the Technology, Health, Oil, and Filmmaking industries, respectively. The percent yields/returns of the model's robo-trading function (green bar), holding the stock share for the full period (red bar #1), and randomly trading without relying on any model predictions (red bar #2) have all been implemented. For the cases of AbbVie and Lionsgate, due to the fact that the percent yield for the automated robo-trading exceeds that of randomly trading and holding the stock, it can be determined that the model was able to capture some sort of effect in order to produce a "beneficial" prediction. Additionally, the percent yield from robo-trading was greater than that of holding the stock, which shows that the model's predictions were somewhat accurate. For the cases of Amazon and Saudi Aramco, due to the fact that the percent yield for the automated robo-trading is less than that of holding the stock for the duration, yet greater than randomly trading, it proves that the model somewhat captured the effect, and exhibited decently accurate predictions; if the model made perfect predictions, holding the stock and automated robo-trading would produce similar percent yields, but this is not the case.

Conclusion from Findings:

Overall, the findings convey that the model was unable to make a prediction that was more accurate for one industry in comparison to the others. In fact, the model was ineffective on that scale and a random implementation of trades would have produced fewer losses and higher returns for every industry; this shows that the model did not capture any relevant effects on an industry-wide scale. However, the model was able to secure positive returns that were higher in comparison to randomly trading and holding shares for the entire duration for a set of outliers; this proves that, perhaps, the model can be improved upon and further implemented in line with financial theory to become more effective.

Future Work

The results obtained in our project explain to us that there is more room for exploration to make an accurate conclusion about the model's efficiency and effectiveness. It was not able to predict more accurately for any industry in comparison with another. However, this was an analysis done on a mere three years of data due to concerns about the effect of COVID-19. For more accurate predictions, we would ideally want to implement more historical data without overfitting the model, which might produce results that provide a greater insight. We were unable to do so with this project because this is computationally expensive and requires more powerful hardware that we did not have access to (to obtain results in a feasible time period). Additionally, our project was centered around the assumption that only one share of a stock was being purchased or sold at a time, which has the potential to produce uncommon or unrealistic results. These obtained results from automated robo-trading could be scaled by a factor of the shares being bought or offloaded, which is unrealistic due to hedging strategies where only a percentage of shares would be purchased or sold based on a specific prediction.

Another pitfall of our project is that, if we had more financial and stock trading knowledge, for example, we would've been able to set up a more effective trading strategy that would've produced results that reflect those seen in industry.

In conclusion, though, this project serves as a baseline foundation for primary experience in the fields of the assessment of regression models, stock share price prediction models, and automated trading strategy (robo-trading). Future improvements to this project would include a more realistic automated trading strategy, a training and testing set that has access to a larger time period of historical data, and a selection of industries and corresponding companies that are more in-line with financial theory (instead of random selection). These future implementations are all realistic and can potentially allow us to gain more accurate knowledge surrounded with our findings. This will therefore allow us to form more relevant conclusions about an ML model's performance in relation to the prediction of something that is, in short, extremely difficult to predict.

Works Cited

Royal, J. (2022, February 4). *Saving vs. investing: Here are the key differences for managing your money*. Bankrate. https://www.bankrate.com/investing/saving-vs-investing/

Goedhart, M., Koller, T., & Wessels, D. (2005, March 1). *Do fundamentals--or emotions--drive the stock market? | McKinsey*. Www.mckinsey.com. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/do-fundamentalsor-emotionsdrive-the-stock-market

Fama, E. F. (1965). Random Walks in Stock Market Prices. *Financial Analysts Journal*, *21*(5), 55–59. http://www.jstor.org/stable/4469865

Avci, R. (2023, March 27). *Challenges With Stock Price Prediction*. Medium. https://medium.com/@rasim.avci/challenges-with-stock-price-prediction-17e151bd79a7#:~:text=However%2C%20there%20are%20several%20challenges

Rational Investing in an Age of Uncertainty | Morgan Stanley. (2023). Morgan Stanley; Morgan Stanley. https://www.morganstanley.com/articles/behavioral-finance

Bora, N. (2021, November 8). *Understanding ARIMA Models for Machine Learning*. Capital One. https://www.capitalone.com/tech/machine-learning/understanding-arima-models/

Speights, K. (2023, November 16). *Most Volatile Stocks Right Now*. The Motley Fool. https://www.fool.com/investing/stock-market/types-of-stocks/volatile-stocks/