Performance of Time Series Analysis in Predicating Volitile Stock Price*

Using ARIMA to Forecast the Stock Price of Gamestop

Mingyang Li

27 April 2022

Abstract

This report builds an ARIMA model of the Gamestop company stock to examine the effectiveness of a statistical approach in equity predictions. Using the time series analysis method, ARIMA(4, 0, 4) is the final model left for forecast. The model is then used to predict 10 daily returns after the testing data cutoff, and the result of the first four was close to the actual data. In conclusion, ARIMA can be used to predict stock performance when the statistical assumption of the model is satisified, but there is also visible boundry to this strategy.

1 Introduction

Time series analysis is a specific statistical practice that is applied to forecast data collected over a constant time interval (Mondal, Shit, and Goswami 2014). The stock market perfectly fits the application of time series analysis because the return of investment can be formulated into a time series, and stock traders often make decisions based on how they believe the company will perform in the future. In fact, the recent trend of stock strategies suggests that more traders have begun to favor technical analysis, which relies on techniques from time series analysis, such as Autoregressive Integrated Moving Average (ARIMA)(Ebert and Hilpert 2019). The motivation behind this trend of technical routines is firstly that it promises fruitful results in investors' portfolios, while other industries have also proven the value of statistics. However, this paper wants to emphasize that statistical models including ARIMA have very high standards for the data to match in order to produce reliable results. As an example, ARIMA models the time series with the presumption that the data is stationary. Also, autoregressive models only work on univariate data. These assumptions imply that the return of a stock should not be dependent on trading days, and its predictability should only depend on the historical returns. Apparently, many people do not agree on assessing an equity by only looking at its price in the past. These fundamental traders believe there is more in stock prediction than its historical stock values. As a result, the objective of this report is to examine whether ARIMA would still be robust when analyzing a stock that has had unexpected performance trends. The company that the report focuses on is GameStop. This company's stock price has been volatile over the past year, and the news about the company's business has been an important catalyst in the stock price. In the remaining sections of this paper, we will first examine and convert the return of Gamestop's stock into a stationary time series. Through a diagnostic check on the models we proposed, we will filter out the best model and compare its prediction with the testing set of data. All of the technical work is conducted through the use of the statistical programming language R (R Core Team 2020).

2 Data

The report uses the data obtained from `tidyquant`, which allows us to get the most up to date stock prices formatted in time series (Dancho and Vaughan 2021). There are 371 observations and 7 variables including

*Code and data are available at: https://github.com/samlmy/gmeARIMA.git index. The raw data contains the opening, closing, daily high, daily low, adjusted price and volume of each regular trading day from November 1, 2020 to April 25, 2022. The variable of interest is the daily closing price, which refers to the price that the stock is traded at by the end of a trading day. When a trading platform shows how much a stock has surged or plummeted in percentage or dollar value, it compares the closing price with the current price to calculate those values. The closing price is also commonly used to portray the trend of a stock. We recreated the plot of stock closing prices on a daily interval using `ggplot2` (Wickham 2016). Figure 1 shows the trend of the GameStop stock price within the past 371 regular trading days. The reason why `tidyquant` is preferred in this study is because it provides the data in a time series format; stock price over regular trading days in this case (Dancho and Vaughan 2021). Furthermore, we constructed a daily return variable for the modeling we needed later. The daily return shows the percentage of increase or decrease of the closing price from current trade day versus the previous day. The purpose of this variable is that modeling ARIMA require the data to be stationary and to have a mean of zero. By plotting the daily return, figure 2 illustrated a time series centered at zero. We assume that this time series is also stationary, because there is no obvious pattern. Combining the daily return to the original dataset provides 370 observations instead of 371. This is because the daily return is the difference of each observation, leaving the 1st of November 2020 to have null value. We took out the last ten days of return as the testing dataset. Tables in this reported is created through `knitr` (Xie 2021). Diagnostics and forecast of ARIMA is done through `astsa` (Stoffer 2021).

Figure 1: GME's daily closing price from November 1, 2020 to April 25, 2022 which is a total of 371 regular trading days

3 Model

Modeling with ARIMA firstly requires us to identify the lag cut offs by plotting the autocorrelation function (ACF) and the partial autocorrelation function (PACF). The first thing to note is that both ACF and PACF

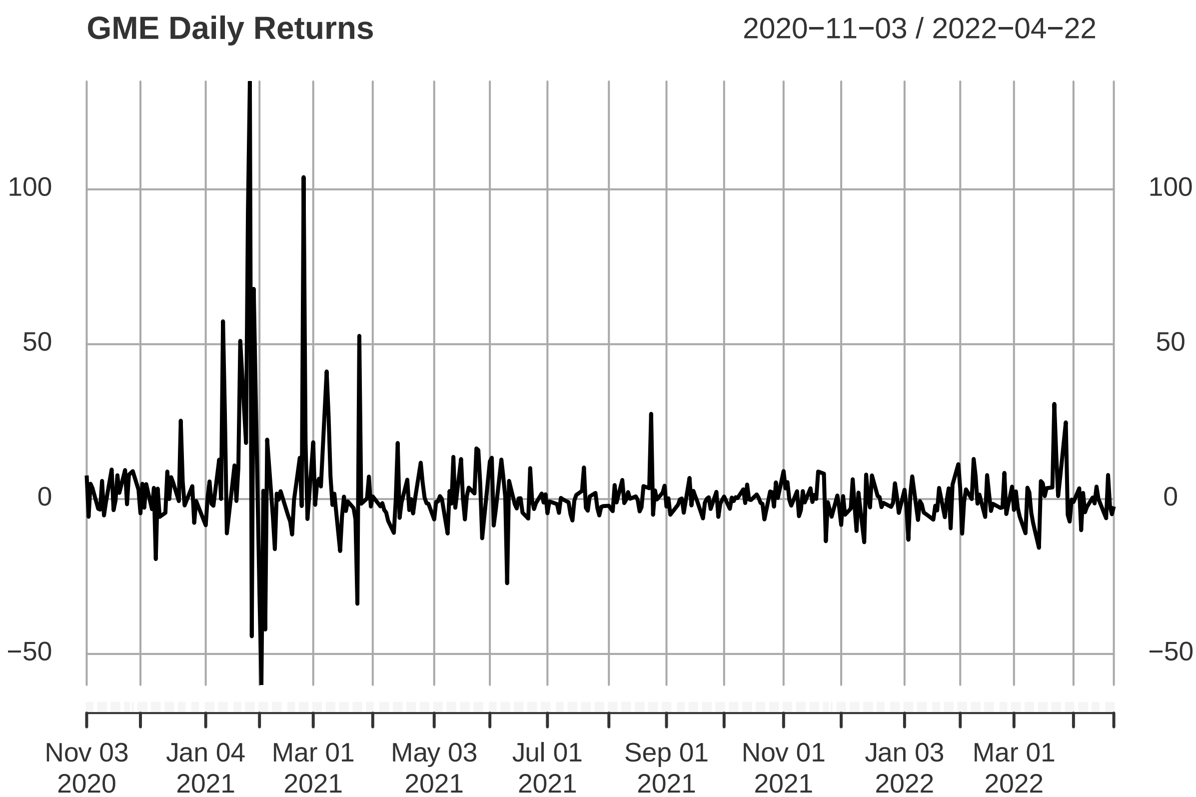


Figure 2: GME's daily return in percentage from November 2, 2020 to April 25, 2022 which is a total of 370 regular trading days

has a fast decay, which means we can use the data as it is and it does not require further transformation. The ACF plot suggests a cut off at lag 4, and PACF also indicates a cut off at lag 4. Hence, we propose the following three models: ARIMA(4, 0, 0), ARIMA(0, 0, 4), ARIMA(4, 0, 4). The selection of model is done by inspecting the Akaike Information Criteria (AIC). AIC correction (AICc) is the value we want to examine in particular, as it tells us whether the model is a good fit when the sample size is finite. Table 1 shows all the corresponding AICc for each model, and the lower the value, the better the fit, which means that ARIMA(4, 0, 4) is on top of the selection. ARIMA(4, 0, 4) is written out as

$$x_t = \mu_0 + \phi_1 \cdot x_{t-1} + \phi_2 \cdot x_{t-2} + \phi_3 \cdot x_{t-3} + \phi_4 \cdot x_{t-4} + w_t - \theta_1 \cdot w_{t-1} - \theta_2 \cdot w_{t-2} - \theta_3 \cdot w_{t-3} - \theta_4 \cdot w_{t-4}$$

x_t is the daily return of gme on tth trading day. w_t is the error term on the tth trading day. μ_0 is the intercept in the daily return that is always added towards the prediction. ϕ_i and θ_i stands for the ith autoregression (AR) coefficient, and moving average (MA) coefficient, respectively. Finally, we need to justify the model by checking whether it meets the assumption of the analysis. This step is done by inspecting the diagnostic plots. We want to see constant variance and mean centered at zero in the Standardized Residuals, meaning there is no dependence in residual errors. There should be no lags surpassing the significance level at 5% in ACF of Residuals, suggesting the residual is random. Normal Q-Q Plot of Std Residuals indicates the assumption of normality when all points fall on the straight line. P-value for Ljung-Box statistic should show points above the 5% significant level to indicate that the residuals are independently distributed. Figure 3 demonstrates the four diagnostics mentioned above, and there are visible issues with the model, such as outliers. However, we can safely say that the assumptions are mostly satisfied for the purpose of this analysis. The parameter estimation of the model will be displayed in the **??** section; more issues about assumptions violation will be further explored in the 5 section.

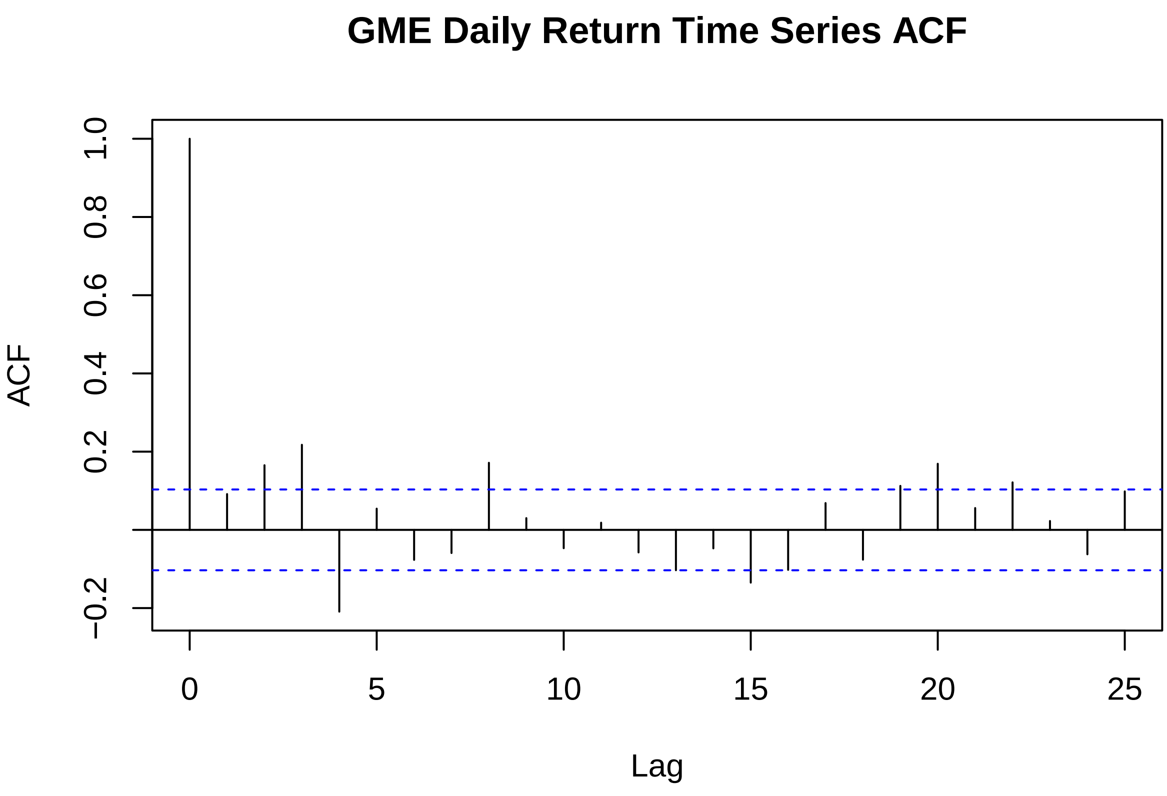
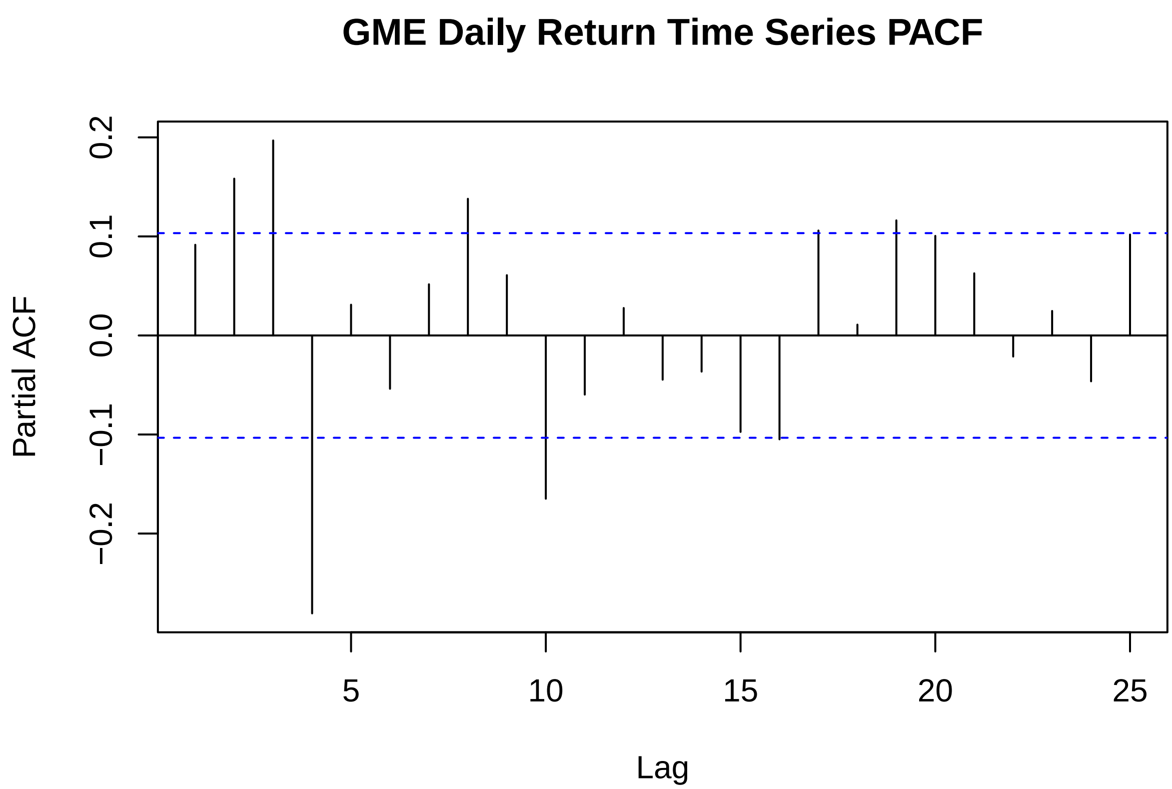


Table 1: AICc of ARIMA(4, 0, 0), (0, 0, 4), (4, 0, 4) in numbered order

	AICc
1	-399.4335
2	-394.6037
3	-401.7029



4 Results

Table 2 shows the estimated parameter of ARIMA(4, 0, 4). The AR coefficients corresponds to ϕ in the equation, and MA are θ. Xmean is the intercept. For example, AR1 means that the predicted return x_t on day t changes by -26.3% of the return on day $t-1$. On the other hand, MA1 suggests that the predicted return on day t will change by 34.22% of the error term on day $t-1$. Adding each term multiplied by the corresponding coefficient will give us the forecast on the given tth trading day. We have left 10 trading days for testing in the preparation phase, thus we will forecast the daily returns of 10 regular trading days after April 8, 2022. The predicted daily return of GME with ARIMA(4, 0, 4) built on training data are shown in figure 4 and table 3. The red data points highlighted in figure 4 stands for the prediction, and the grey area is the 95% confidence interval of the forecast return. The actual daily returns are displayed in table 4 It turns out that the first five predicted returns closely resembles the actual data, with a difference less than 1.7%. More importantly, the model successfully predicted a gain or loss for the first five trading days. The difference between the forecast and actual return are indeed within 0.5%, which is exceptionally accurate. However, it is noteworthy that the five subsequent days of prediction parted away from the actual returns by quite a lot. Not only was the direction of return opposite, but some of the values were also off by more than 7%. Overall, we conclude that ARIMA would be suitable for stock analysis, even if the equity is generally unstable and extremely volatile.

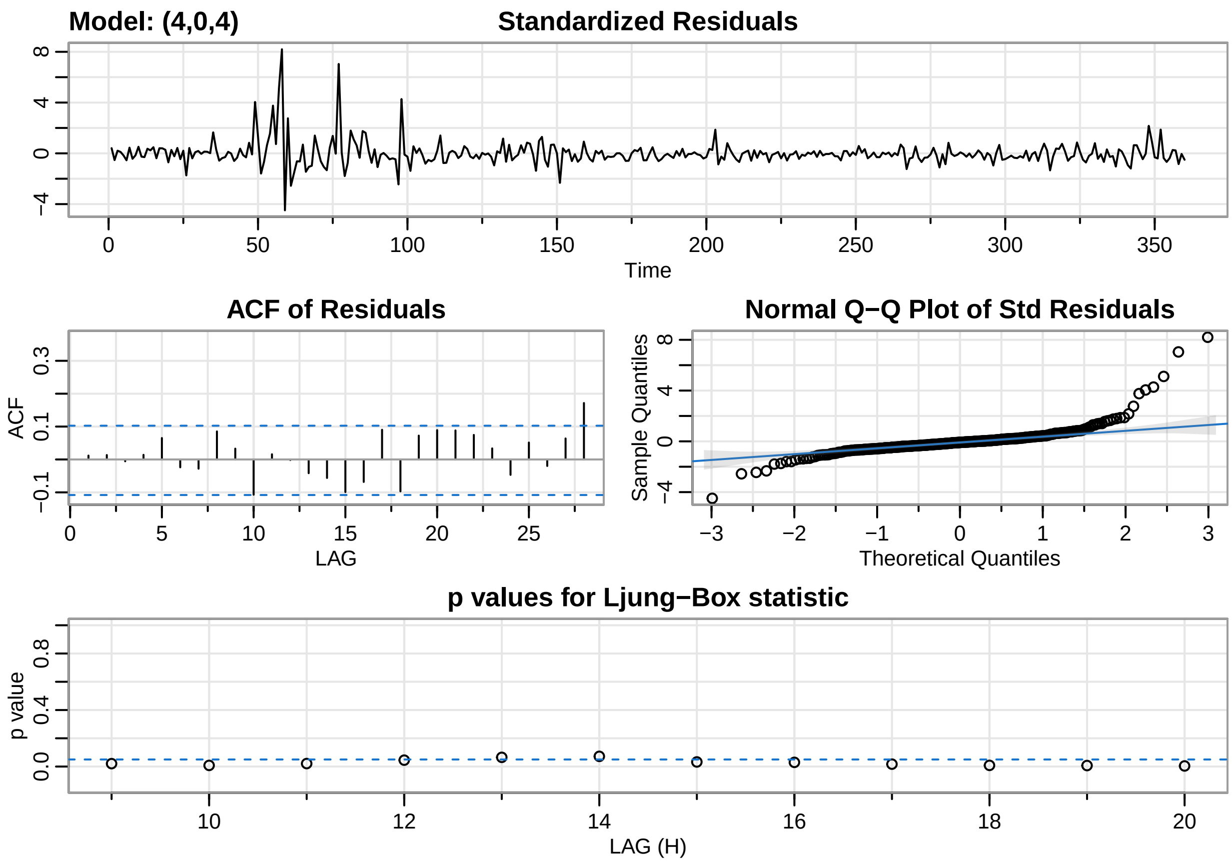


Figure 3: ARIMA(4, 0, 4) assumption check diagnositics

Table 2: Intercept and Coefficient of ARIMA(4, 0, 4)

	Coefficient
ar1	-0.2630
ar2	-0.0152
ar3	-0.1687
ar4	-0.4272
ma1	0.3466
ma2	0.2131
ma3	0.5185
ma4	0.3338
xmean	0.0161

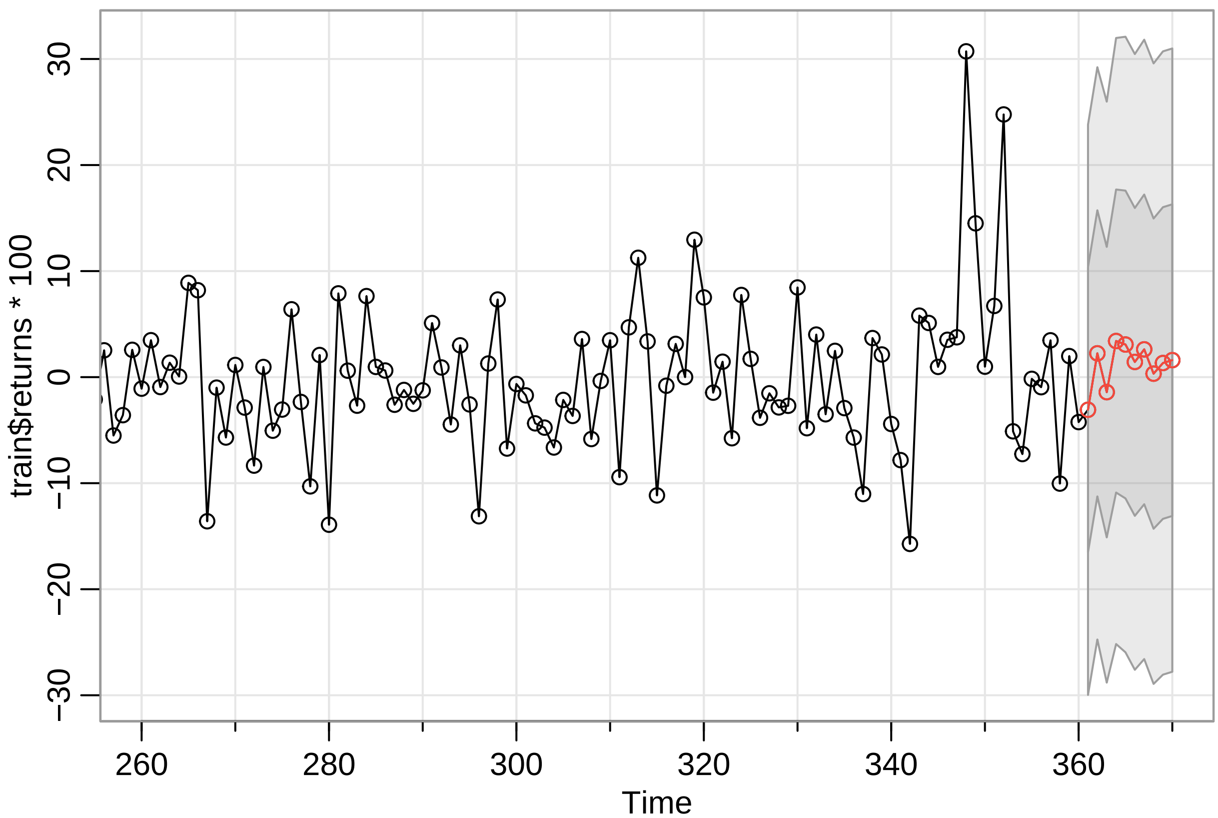


Figure 4: GME Predicted Daily Returns with ARIMA(4, 0, 4) in Percentage

Table 3: Predicted Daily Returns of ARIMA(4, 0, 4)

	returns
1	-3.0784905
2	2.2434685
3	-1.4159106
4	3.4073037
5	3.0750159
6	1.4326360
7	2.6193974
8	0.3278083
9	1.3314550
10	1.6038360

Table 4: Actual Daily Returns of GME

returns
-2.53999867
0.50618441
-1.40202688
4.02429917
0.04645455
-6.17496634
7.78311428
-2.37423426
-4.89755304
-2.35942148

5 Discussion

5.1 Assumptions Violation

The assumptions for statistical analysis are very important to modeling because different models are designed to fit specific types of data. While in the real world, no data can satisfy the model assumptions perfectly, the conditions should match as closely as possible for the model to work as it is intended. In earlier sections, we mentioned that ARIMA works on stationary time series, which means that the data are independent. To check in detail, statistical values such as residuals are computed and plotted to examine whether they have matched the criteria. We saw that most p-values for Ljung Box statistic shown in figure **??**fig:4) are not ideal. Even though the model had most predictions matched with the actual data, it still suggests ARIMA might not be a good fit for stock like GME or stocks in genergal.

5.2 Limitation of ARIMA prediction

In the 3 section, the written equation of ARIMA(4, 0, 4) clearly indicates that the model relies on the return and error term of the past four trading days. Thus, if the range of prediction exceeds four days, the forecast is indeed computed on predicted values rather than the actual historical data. This is reflected in our results where the first four predictions are much more accurate than the other ones, and the reason why the results begin to part off is because the model is relying on fabricated values.

5.3 Relation between individual stock and the stock market

It should be noted that volume is a significant factor when it comes to the technical analysis of stocks. Using what we can find on data and news, the extreme values always happen on a trading day with an enormous amount activity, reflected in volume. During regular trading days, if the particular stock does not have catalyst news that serves to boost or tank the stock price, the trend of the stock particularly yields to the trend of the index or the ETF that it belongs to. With that being said, using ARIMA during times when the company does not have any special business movements would be more reliable, since ARIMA is univariate. On the other hand, this strategy should be avoided when the stock market is filled with speculation and uncertainty.

5.4 Weaknesses and next steps

One of the weaknesses of this report is that it only focuses on one particular stock. Even though it might be the most representative out of the most volatile equities, more cases would strengthen the findings of this report.

References

Dancho, Matt, and Davis Vaughan. 2021. *Tidyquant: Tidy Quantitative Financial Analysis*. https://github.com/business-science/tidyquant.

Ebert, Sebastian, and Christian Hilpert. 2019. "Skewness Preference and the Popularity of Technical Analysis." *Journal of Banking & Finance* 109. https://doi.org/10.1016/j.jbankfin.2019.105675.

Mondal, Prapanna, Labani Shit, and Saptarsi Goswami. 2014. "STUDY OF EFFECTIVENESS OF TIME SERIES MODELING (ARIMA) IN FORECASTING STOCK PRICES." *International Journal of Computer Science, Engineering and Applications* 4 (2): 13. https://doi.org/10.5121/ijcsea.2014.4202.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stoffer, David. 2021. *Astsa: Applied Statistical Time Series Analysis*.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.