Forecasting Stock Market Prices Using ARIMA

Nicholas D'Agostino
Pace University

***ABSTRACT* – This research paper investigates the use of the Autoregressive Integrated Moving Average (ARIMA) model in predicting stock prices. Employing historical stock data retrieved through the yfinance library, the study examines the effectiveness of ARIMA in discerning dependencies within financial time series. My contribution lies in the examination of ARIMA's potential as a forecasting tool, grounded in rigorous evaluation metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared.**

I. INTRODUCTION

Financial markets pose challenges due to their dynamic and unpredictable nature, demanding sophisticated decision-making tools. Forecasting stock prices is critical for informed investment decisions and risk management. Amid various methods available for stock price prediction, this study focuses on the Autoregressive Integrated Moving Average (ARIMA) model. The methodology encompasses data retrieval, preprocessing, iterative differencing, model fitting, and comprehensive evaluation metrics. Subsequent sections will expand on the literature review, methodology, results, discussion, and conclusion.

II. LITERATURE REVIEW

The field of stock price prediction has evolved with diverse methodologies, encompassing traditional approaches like moving averages and autoregressive models. Recent emphasis on machine learning algorithms, despite their popularity, is accompanied by challenges such as overfitting and data sensitivity. Many studies, including this one, aim to contribute by scrutinizing the ARIMA model, a robust time series forecasting technique, addressing potential gaps in predictive modeling within finance.

In 2019, Peter T. Yamak et al. conducted a comprehensive study comparing the performance of different models, including ARIMA, LTSM, and GRU, in the context of stock price prediction [6]. Their analysis delved into the intricacies of the ARIMA model, which is dissected into three key components: AR for autoregression, I for integrated differencing, and MA for moving average. The study provided a mathematical decomposition of the ARIMA model equation, highlighting the fusion of a pure autoregressive model with a moving average model to formulate the complete ARIMA equation. Additional insights were provided regarding the relevant parameters (p, d, and q), with emphasis on addressing trends, seasonality, and stationarity.

Subsequent sections of the study explored the LSTM and GRU models, offering background information on their methodologies. The researchers proceeded with data collection and processing, utilizing Bitcoin's transactional data spanning from 2014 to 2019. The normalization and transformation of the data were integral steps, especially considering the utilization of the ARIMA model, which necessitates testing for stationarity. This validation was achieved through the Augmented Dickey Fuller (ADF) test, wherein the null and alternative hypotheses were declared, and significance was measured via the interpretation of the p-value. A p-value greater than 0.05 indicated a lack of substantial evidence to reject the null hypothesis, implying non-stationarity, whereas a p-value less than 0.05 signaled stationarity.

Importantly, the same ADF test methodology is later employed in my research, as detailed ahead. Following the rigorous testing and fitting of all models, the study concluded that the ARIMA model outperformed its counterparts, LSTM and GRU, in terms of both accuracy and computational efficiency. This determination was based on lower Mean Absolute Percentage Error (MAPE) and

Root Mean Squared Error (RMSE) scores for the ARIMA model.

In a study conducted in 2021, Sampat Kumar et al. developed a predictive model using cryptocurrency closing prices [8]. They provided a clear and intuitive breakdown of the fundamental structure of the Autoregressive Integrated Moving Average (ARIMA) model, elucidating its components. Following this foundational understanding, the researchers outlined their methodology, which involved the collection of data from Yahoo Finance. Filtering and cleaning procedures were implemented to address data irregularities, such as null values, and remove irrelevant features. During the preprocessing stage, the researchers employed techniques such as feature engineering and the use of rolling windows to enhance the depth of their analysis. The visual representation of their methodology paired with their feature engineering offered valuable insights, influencing the structuring of my own approach in this research. After constructing and fitting the ARIMA model to its parameters, an evaluation was conducted. The performance metrics indicated minimal forecast error, showcasing a nearly perfect fit, as evidenced by an R-Squared value of 0.991.

Overall, these studies, with the addition of several others to supplement, served as an invaluable guide, particularly for individuals who are new to the field of time-series forecasting.

III. METHODOLOGY

A. Data Collection

Historical stock data for AAPL was acquired using the yfinance library, covering the period from 2015 through 2020. The dataset includes daily open, close, high, and low prices; however, the closing price will be treated as the pivotal variable for our forecasting model.

B. Data Preprocessing

Primary preprocessing steps involved the handling of missing values and converting the index to a business day frequency. These procedures are fundamental for ensuring data quality and consistency. Additional preparations are made in the form of visualizing data and testing/removing stationarity. To validate stationarity, make use of the Augmented Dicky Fuller (ADF) test [6]. Visualizations and testing are displayed in the following figures.

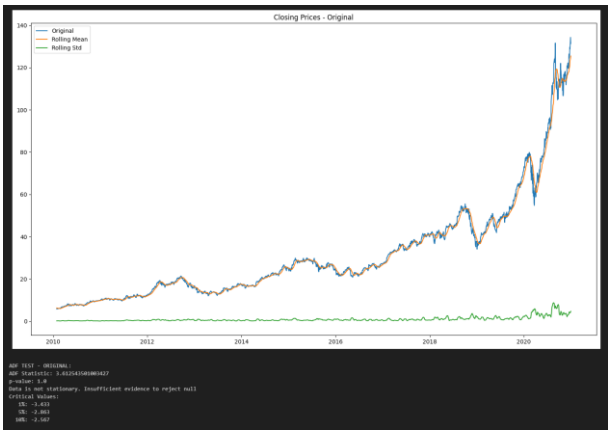

Figure 1: Graph for ADF test prior to differencing

Recall that the ADF test is meant to return a p-value which is then used to either confirm or reject the hypothesis given. The hypotheses are as follows:

H_0: The data is not stationary

H_i: The data is stationary

Confirm these hypotheses under the following interpretation:

p-value > 0.05: Accept H_0. Data is not stationary

p-value < 0.05: Accept H_0. Data is stationary

Upon close inspection of the ADF test results displayed in Figure 1, observe that the p-value is 1.0, thus there is insufficient evidence to reject the null hypothesis and suggest that the data is not stationary.

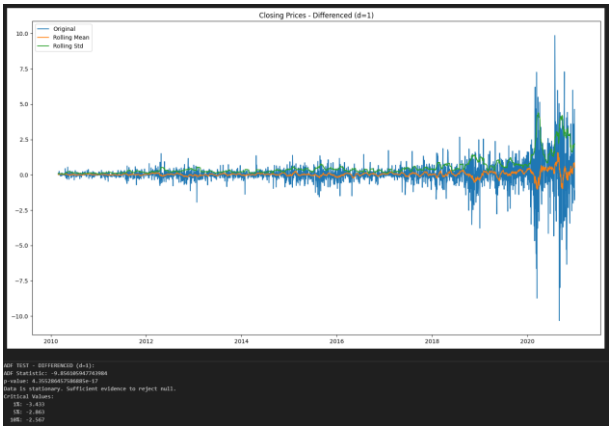

Figure 2: Graph for ADF test after differencing

Upon close inspection of the ADF test results displayed in Figure 2, observe that the p-value is 4.355286457586885e-17, thus there is sufficient evidence to reject the null hypothesis and suggest that the data is stationary. Now, it can be concluded that one order of differencing (d=1) is required to make the data stationary and ready for fitting.

C. ARIMA Model Implementation

The ARIMA model is implemented with iterative differencing to attain stationarity [4]. Parameters p, d, and q are carefully selected to capture underlying patterns in the data where:

p: The autoregressive (AR) order

d: The differencing (I) order

q: The moving average (MA) order

The ARIMA (p, d, q) model equation is defined as follows:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_q \epsilon_{t-q}$$

where *Yt* is the observed time series value at time *t*, α is the mean of the series, β_p are the autoregressive coefficients, ϵ_t are the white noise error terms, and ϕ_q are the moving average coefficients [6, 7].

As was already determined, to achieve stationarity, the differencing order, or d, must be at least 1. To obtain the AR(p) and MA(q) parameters, fit an ARIMA (0, 0, 0) model. Using the residuals, plot the long/short-term autocorrelation and partial autocorrelation functions for the newly fitted model. Observing decay in both the ACF and PACF plots will indicate what the AR(p) and MA(q) parameters will be [1].

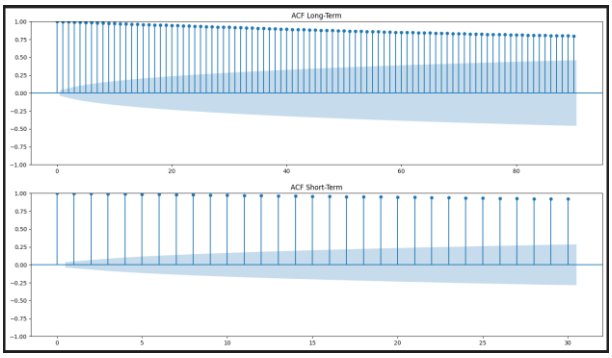

Figure 3: Graph for Long/Short-term ACF

In Figure 3, identify the gradual decay in correlation values over time. This observation is indicative of heavy dependence between future values and lagged values, suggesting that there is a moving average, or MA, of 1 [5].

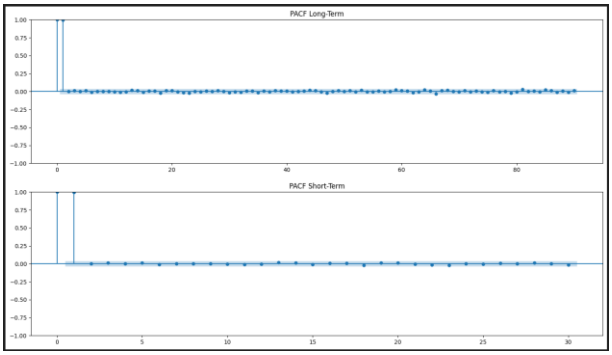

Figure 4: Graph for Long/Short-term PACF

In Figure 4, the sharp spike and decay in lag values at a lag of 1 indicates there be an autoregression, or AR, order of 1 as well. These observations have led to the belief that the best fit model would be ARIMA (1, 1, 1) [5].

IV. RESULTS

The implementation of the ARIMA model provides valuable insights into forecasting accuracy through visualizations (figure 5)

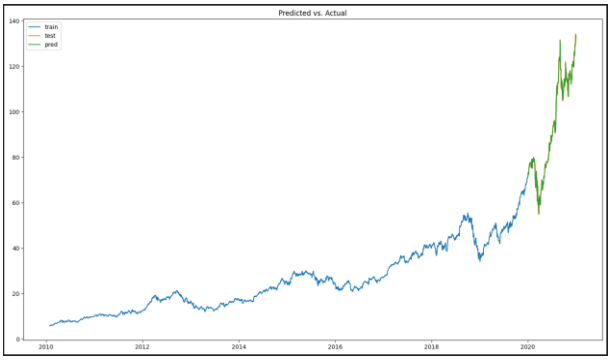

Figure 5: Graph of Actual vs Predicted Values

and performance metrics. Metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) are employed to evaluate the model performance against actual stock prices as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where n is the number of observations, Yi is the actual value at time i, $\hat{Y}$ is the predicted value at time i, and $\bar{Y}$ is the mean of the actual values [2, 3].

```
Mean Squared Error (MSE): 6.2533
Root Mean Squared Error (RMSE): 2.5007
Mean Absolute Error (MAE): 1.7588
R-squared (R2): 0.9864
```

Figure 6: Performance Metrics

V. DISCUSSION

The results obtained from the implementation of the ARIMA model on historical AAPL stock data demonstrate promising forecasting accuracy, as evidenced by the visual comparison of actual vs predicted values (Figure 5) and the comprehensive evaluation metrics (Figure 6). The choice of performance metrics, including MSE, RMSE, MAE, and R-squared, ensures a thorough and balanced assessment of the model's predictive capabilities.

The initial non-stationarity of the data was addressed through first-order differencing, as confirmed by the ADF test results (Figure 2). This step is essential for the effectiveness of the ARIMA model, as it requires stationary time series data for accurate forecasting.

The determination of the ARIMA model parameters (p, d, q) involved careful analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots (Figure 3 and Figure 4). The observed patterns in these plots led to the selection of an ARIMA (1, 1, 1) model, suggesting a combination of autoregressive and moving average components with differencing to achieve stationarity.

The visual representation of actual vs predicted values (Figure 5) indicates a close fit between the model predictions and observed stock prices. Additionally, the performance metrics (Figure 6) further confirm the accuracy of the ARIMA model, with low values for MSE, RMSE, and MAE, and a high R-squared value. These metrics collectively showcase the model's ability to capture the temporal dependencies within the financial time series.

VI. CONCLUSION

This research paper contributes to the ongoing exploration of forecasting stock prices using the ARIMA model. The study demonstrates the effectiveness of ARIMA in capturing temporal dependencies within financial time series, particularly in the context of AAPL stock data. The meticulous methodology, including data collection, preprocessing, and parameter selection, ensures the reliability of the results.

The findings of this research have implications for financial analysts and investors, providing them with a tool for informed decision-making and risk management. The ARIMA model, with its ability to discern and predict patterns in stock prices, can serve as a valuable addition to the arsenal of forecasting techniques in the financial domain. However, it is important to note that while ARIMA shows promise, no forecasting model is infallible, and market dynamics are subject to various external factors.

REFERENCES

[1] Capital One. (2023, June 8). "ARIMA Model Tips for Time Series Forecasting." https://www.capitalone.com/tech/machine-learning/arima-model-time-series-forecasting/

[2] Fernando, J. (n.d.). “R-squared: Definition, calculation formula, uses, and limitations.” Investopedia. https://www.investopedia.com/terms/r/r-squared.asp

[3] G. W. R. I. Wijesinghe and R. M. K. T. Rathnayaka. (2020). "Stock Market Price Forecasting using ARIMA vs ANN; A Case study from CSE." In 2020 2nd International Conference on Advancements in Computing (ICAC), Malabe, Sri Lanka, 269-274. https://doi.org/10.1109/ICAC51239.2020.9357288

[4] Hayes, A. (n.d.). "Autoregressive Integrated Moving Average (ARIMA) Prediction Model." Investopedia. https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

[5] Monigatti, L. (2023, December 4). “Interpreting ACF and PACF plots for time series forecasting.” Medium. https://towardsdatascience.com/interpreting-acf-and-pacf-plots-for-time-series-forecasting-af0d6db4061c

[6] Peter T. Yamak, Li Yujian, and Pius K. Gadosey. (2020). "A Comparison between ARIMA, LSTM, and GRU for Time Series Forecasting." In Proceedings of the 2019 2nd International Conference on Algorithms, Computing, and Artificial Intelligence (ACAI '19), 49–55. Association for Computing Machinery, New York, NY, USA. https://doi-org.rlib.pace.edu/10.1145/3377713.3377722

[7] Prabhakaran, S. (2023, September 8). "ARIMA Model - Complete Guide to Time Series Forecasting in Python: ML+." Machine Learning Plus. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

[8] S. K. U, S. P. Aanandhi, S. P. Akhilaa, V. Vardarajan, and M. Sathiyanarayanan. (2021). "Cryptocurrency Price Prediction using Time Series Forecasting (ARIMA)." In 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 598-602. https://doi.org/10.1109/ISRITI54043.2021.9702842