Comparative Analysis of Linear Regression and ARIMA Models for Stock Price Forecasting

Boya Huang

Milton International School, Qingdao, China

sodayellow2025@outlook.com

Abstract. Stock price forecasting has long been a critical area in finance, as accurate predictions provide valuable guidance for investors, portfolio managers, and policymakers in navigating uncertain markets. Traditional statistical models remain widely used because of their interpretability and relatively low computational cost, even as more complex machine learning methods gain popularity. This study investigates the effectiveness of linear regression and Autoregressive Integrated Moving Average (ARIMA) models in forecasting stock prices. Using historical data from APPL, GOOG, TSLA, NVDA and MSFT, the models were evaluated under different market conditions. Results show that linear regression provides strong in-sample fitting, especially for stable stocks with lower volatility. However, it struggles with highly volatile stocks due to sudden market shocks. In contrast, ARIMA captures time-series dynamics more effectively and delivers more accurate short-term forecasts, especially in volatile environments. The findings suggest that both models have complementary strengths, and hybrid approaches may further improve stock price prediction in complex financial markets.

Keywords: Stock price forecasting, Financial modeling, Linear regression, ARIMA.

1. Introduction

Recently, stocks have increasingly attracted people's attention. Stocks are frequently brought on by social and political policies. However, it is difficult for investors to make money in the stock market because of the intrinsic volatility of stock values. By this way, accurate stock price forecasting has become a critical focus for investors, analysts and policymakers. The value of stocks fluctuates in a seemingly random manner, yet it can be estimated using statistical models. Among these, regressive models, particularly the Autoregressive Integrated Moving Average (ARIMA) framework and linear regression, can be considered to find patterns among data. Therefore, this study aims to investigate the viability and limitations of regressive models in forecasting stock prices, drawing on both academic literature and empirical experimentation.

In the last decade, researchers have used many prediction models on stock data. Zhao takes BYD as an example, the performance of the multiple linear regression model and the BP neural network in stock price prediction was compared. It was found that there was a significant correlation between the stock price and the exchange rate (USD/CNH), CPI, P/E ratio, and Dow Jones Index, and the BP neural network had better explanatory power than the multiple regression model [1]. Hu uses Python to build a multivariate linear regression model, they used Alphabet as an example to predict stock prices. The results showed that the difference between the predicted and true values was small, with high accuracy [2].

ARIMA is a statistical model for time series data that captures stationarity and linear relationships [3]. Sunki et al. compared three models and found ARIMA works well when the underlying data is steady and linear [3]. According to the findings from the Haider et al, the ARIMA model could be useful for financial managers and investors when they're making decisions about Meta Stocks. The results show that the ARIMA model can work well with traditional stock price predictions. They also emphasize that Meta Stocks has a lot of potential for short-term forecasts, even with its very unpredictable price swings [4]. As of 2025, global stock markets face increased volatility. Rising interest rates, inflation, and geopolitical risk have created more frequent price shocks. Markets show less stability, and regime shifts are common. Despite this, there is an important research gap. A few

studies compare linear regression and ARIMA across different types of market conditions to carry out further investigation.

The aim of this research is to compare the effectiveness of linear regression and ARIMA models in forecasting stock prices, and evaluate which model performs better under different market conditions and to identify the strengths and limitations of each method. This study will improve understanding of how classical statistical models can remain useful in today's complex financial markets. To conduct this research, historical stock price data from publicly listed companies will be collected. Stocks from different industries will be selected to test model performance under varying conditions. Both ARIMA and linear regression models will be built using Python, and their performance will be evaluated using standard metrics.

After testing the models, results will be compared to determine which model fits each type of data more effectively. Situations in which each model becomes unreliable or overly simplistic will also be discussed. Finally, improvements or hybrid approaches will be suggested based on the findings. This study intends to provide practical insights for investors, while also highlighting areas where traditional models require updates or replacement.

2. Method

2.1. Dataset Preparation

The dataset used in this study comes from Kaggle [5]. It includes daily price records of five public companies: Apple (AAPL), NVIDIA (NVDA), Microsoft (MSFT), Tesla (TSLA), and Alphabet (GOOG). The time range of the data is from April 7, 2009, 05:00:00+00:00 to July 4, 2025, 04:00:00+00:00.

Each stock has 15,161 data points. There are 75 variables for each record. This paper uses four of them: stock_close, stock_volume stock_high and stock_low. he trend of this data can be found in Fig. 1.

Fig. 1 Closing Price Trend of AAPL, GOOG, TSLA, MSFT, NVDA (Picture credit: Original)

2.2. Linear regression

$$stock\_low = \beta_0 + \beta_1 \times stock\_volume + \beta_2 \times stock\_high + \beta_3 \times stock\_low + u \quad (1)$$

Stock_close represents daily closing price, stock_high is daily high price, stock_low is daily low price and stock_volume means trading volume. The equation can be shown below:

Multiple regression is a way to study how one thing changes when many other things change [6-8]. It uses more than one input to predict an output. Among these, "stock_close" is chosen as the prediction target. The Close price is often used to show the final value of a stock on a trading day. It is also commonly used in financial modeling and analysis. In this equation, stock_close is dependent variable, stock_volume stock_high and stock_low are independent variables. Stock_close is daily closing price, stock_high is daily high price, stock_low is daily low price and stock_volume is trading volume

Multiple regression has many good points. It can study more than one factor at the same time. This helps to understand how each factor affects the result. Multiple regression also shows relationships clearly. It also helps to predict what may happen. For example, if a company knows what factors affect sales, it can change them to sell more. That can help planning and making better choices. However, multiple regression also has some problems. First, it needs a lot of data. If the data is not correct, the result will be wrong. Second, it can be hard to understand. Math is more complex than simple regression. Not everyone can use it easily. Also, if some of the input data is very similar, the model may not work well. This is called multicollinearity. It makes the result unclear. Sometimes, the model fits the data too well and does not work for new data. This is called overfitting.

2.3. Arima Model

For the company's stock_close, suppose to use the ARIMA (1,1,1) model:

$$(1 - \varphi L)(1 - L)stock_{close_t} = c + (1 + \theta L)\varepsilon_t \quad (2)$$

After expansion:

$$stock\_close_t - stock\_close_{t-1} = c + \varphi(stock\_close_{t-1} - stock\_close_{t-2}) + \varepsilon_t + \theta\varepsilon_{t-1} \quad (3)$$

$$\Delta daily\_close_t = c + \varphi\Delta daily\_close_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1} \quad (4)$$

For these coefficients, first of all, φ (phi) is the AR (1) coefficient, indicating the impact of the previous period's price on the current price. Then θ (theta) is the MA (1) coefficient, representing the influence of the error in the previous period on the current price. c is a constant term, and d=1 indicates that the experiment uses first-order difference to make the sequence stationary

ARIMA is a model used to predict future data. It looks at past values and patterns. ARIMA has three parts: Auto-Regressive (AR), Integrated (I), and Moving Average (MA). It works well for time series data. People use it to forecast sales, prices, and many other things [9, 10].

For advantages, firstly, ARIMA can make good predictions. It uses past data to find patterns. Then, it uses these patterns to guess future values. This helps with planning and making better choices. Second, ARIMA works well with time series data. If the data has trends or regular changes, ARIMA can catch them. It can also remove noise and make the trend clear. However, it might be some drawbacks. ARIMA is not good at handling sudden events. Furthermore, ARIMA can be hard to build. Also, ARIMA works best for short-term forecasting. If the data is used to predict too far in the future, the results may be wrong.

3. Results and Discussion

As shown in Table 1, The R^2 of all models train on different companies is higher than 0.998, indicating that the models can almost perfectly explain the fluctuations of closing prices. Among them, the models of Apple and Microsoft performed the best, with an R^2 of 0.9999, indicating that their closing prices can almost entirely be determined by the highest price, the lowest price, and trading volume.

Table 1. Coefficients of regression

company	R-squared	β_1	β_2	β_3
APPL	0.9999	-4.06e-11	0.5172	0.4830
GOOG	0.9998	-1.79e-10	0.4168	0.5849
MSFT	0.9999	-1.79e-09	0.4911	0.5095
TSLA	0.9997	3.43e-12	0.4086	0.5958
NVDA	0.9995	1.61e-09	0.4540	0.5484

Among the four companies except Apple, it was significant ($P < 0.05$), especially the P value of the stock_volume for NVDA and TSLA was much less than 0.001, indicating that trading volume plays an important role in predicting their closing prices. The stock_volume coefficient of APPL was not significant ($P = 0.301$), indicating that the marginal impact of its trading volume on the closing price was not strong. The magnitude of the coefficient indicates the impact of each unit change in the independent variable on the closing price. For instance, in MSFT's model, the coefficient of stock_high is 0.491 and that of stock_low is 0.508, indicating that the daily price range has a linear impact on the closing price and the weights are close.

The model performs better on low-volatility stocks. For instance, for Apple and Microsoft, the regression lines are close to the actual trends, indicating that linear models can fit stable trends very well. However, for highly volatile stocks, the error rate significantly increases, such as Nvidia and Tesla. The models have difficulty capturing the sharp fluctuations in prices, especially when the market conditions change suddenly, and the predictions are lagging and biased. Linear models are applicable to the prediction of stock closing prices, especially under short-term high-frequency data, where the highest and lowest prices have extremely strong explanatory power for the closing price. Firstly, the predictive effect of trading volume varies among companies. For instance, it is more crucial in TSLA and MSFT but has less impact in APPL. Moreover, the R-square of each model is extremely high, indicating that this method has strong fitting ability within samples. However, attention still needs to be paid to its adaptability to external samples or different market environments.

Overall, it is significant for all companies ($P < 0.001$) and has a relatively high t-value, indicating that the stock price range has a strong explanatory power for the closing price. To explain the coefficients, the influence directions of stock_high and stock_low on stock_close are consistent.

Table 2. Coefficients of ARIMA

company	c	φ	θ
APPL	0.0420	0.9877	-0.9923
GOOG	0.0332	0.8614	-0.8856
MSFT	0.0936	0.2781	-0.3417
TSLA	0.0751	0.9972	-1
NVDA	0.0305	-0.2950	0.1911

Through the ARIMA model shown in Table 2, for APPL, GOOG, and TSLA, the AR coefficients are very close to 1. This means that their stock price changes are highly persistent. A shock or movement in the market tends to carry on for some time. The MA terms for these companies are strongly negative, also close to -1. This suggests that new shocks are often followed by an opposite movement, as if the market quickly corrects itself after an extreme change. These results show that for large technology companies like APPL and GOOG, as well as a more volatile company like TSLA, recent market movements are very important for predicting the near future. However, because shocks last for a long time, these stocks are also harder to predict over longer horizons.

MSFT is different from the first group. Its AR coefficient is much lower, and the MA term is also smaller. This means that Microsoft's stock price changes are less dependent on past values, and the effect of shocks is weaker compared with APPL, GOOG, and TSLA. The dynamics are more stable, and the short-term correction effect is not as strong.

NVDA shows another unique case. The AR coefficient is negative, while the MA coefficient is small but positive. A negative AR term suggests mean reversion: when prices go up, the next movement is more likely to go down, and when prices fall, the next movement is more likely to rise. The positive MA term shows that shocks may sometimes be reinforced instead of corrected. This behavior makes NVDA different from the other companies, since its price movements may not follow the same persistence pattern.

Overall, the results show that large technology stocks do not move in exactly the same way. APPL, GOOG, and TSLA are highly persistent and shock-driven, MSFT is more stable, and NVDA shows some mean-reverting behavior. These differences suggest that investors cannot use the same prediction strategy for all companies. While persistence in some stocks makes them sensitive to recent shocks, mean reversion in others offers a different kind of predictability. The findings also show that ARIMA models can capture these differences and provide useful insights for short-term forecasting.

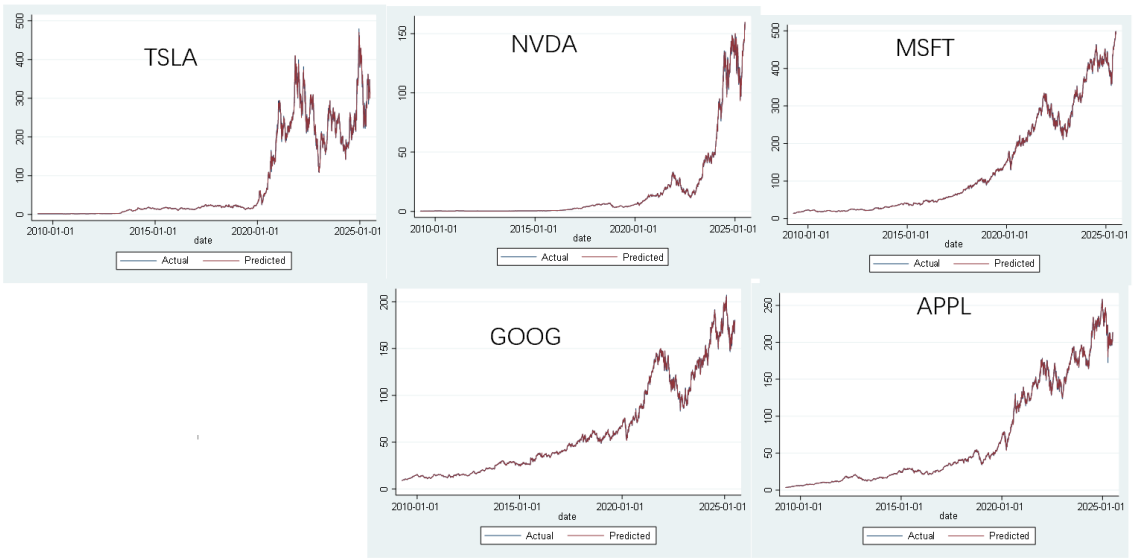

Fig. 2 Prediction results of regressions (Picture credit: Original)

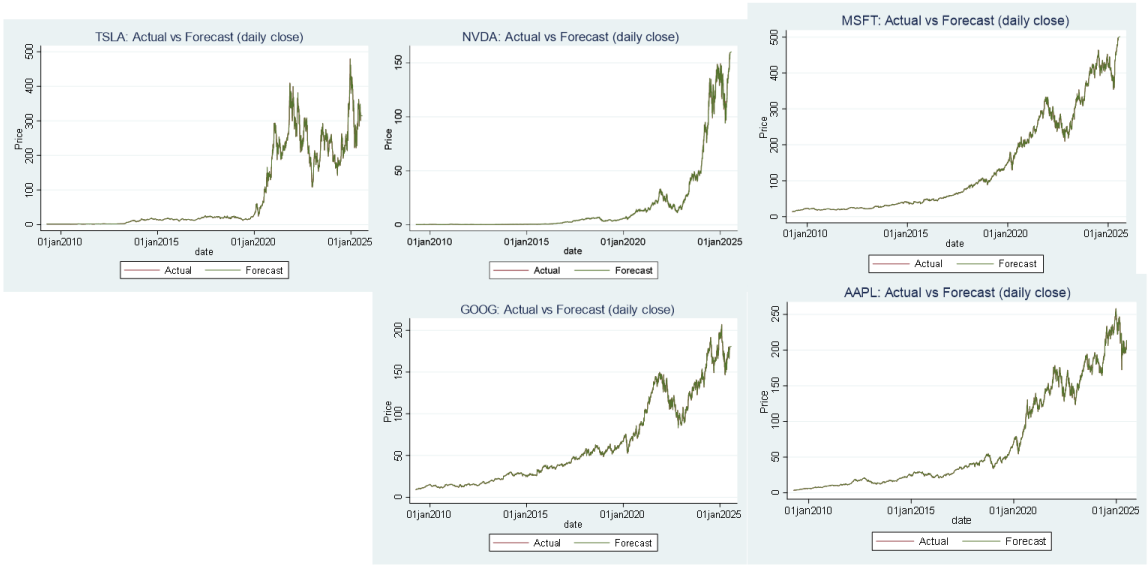

Fig. 3 Prediction results of ARIMA (Picture credit: Original)

Looking at Fig. 2 and Fig. 3, overall, by comparing the prediction figures of regression and ARIMA models, it is evident that ARIMA achieves more accurate forecasts. Although regression provides excellent in-sample fitting, it struggles with highly volatile stocks where market conditions change rapidly. ARIMA, on the other hand, captures the time-series properties of stock data more effectively, producing forecasts that are closer to actual prices, especially in the short term. Therefore, ARIMA is a more suitable approach for modeling stock closing prices when dealing with dynamic and volatile markets.

4. Conclusion

This study started by asking whether linear regression or ARIMA can help forecast stock prices more accurately. The findings show that both models have value but perform best under different conditions. Linear regression works very well when stock prices move steadily. It also shows that stock_high, stock_low, and stock_volume can affect the final price of stocks.

The main contribution of this paper is the comparison of these two models using real data from five major companies. The results demonstrate where each method performs best and where it fails. A key limitation is that both models assume the future resembles the past, which is not always the case in fast-changing markets. In future research, further studies are planned to test hybrid models that combine regression with machine learning in order to better handle large market changes. Such approaches could improve forecast accuracy in real markets.

References

[1] Zhao Y. A comparative analysis of multiple linear regression models and neural networks for stock price prediction—take BYD as an example. In: 2022 2nd International Conference on Enterprise Management and Economic Development (ICEMED 2022). Atlantis Press; 2022. p. 221-6.

[2] Hu R. Stock price prediction based on multiple linear regression model. In: 2023 International Conference on Finance, Trade and Business Management (FTBM 2023). Atlantis Press; 2023. p. 439-47.

[3] Sunki A, et al. Time series forecasting of stock market using ARIMA, LSTM and FB prophet. MATEC Web Conf. 2024;392 (5):01163.

[4] Haider G, et al. Forecasting stock prices: Exploring the potential of ARIMA model for short-term predictions. Int J Manag Res Emerg Sci. 2024;14 (4).

[5] sinankr. tech_stocks_dataset [dataset]. Kaggle; 2009-2025. Available from: https://www.kaggle.com/datasets/sinankr/tech-stocks-dataset.

[6] Sen A, Srivastava M. Multiple regression. In: Regression Analysis: Theory, Methods and Applications. Berlin: Springer; 1990. p. 28-59.

[7] Jeon EH. Multiple regression. In: Advancing Quantitative Methods in Second Language Research. Routledge; 2015. p. 131-58.

[8] Kelley K, Maxwell SE. Multiple regression. In: The Reviewer's Guide to Quantitative Methods in the Social Sciences. Routledge; 2018. p. 313-30.

[9] Shumway RH, Stoffer DS. ARIMA models. In: Time Series Analysis and Its Applications: With R Examples. Cham: Springer; 2017. p. 75-163.

[10] Newbold P. ARIMA model building and the time series analysis approach to forecasting. J Forecast. 1983;2 (1):23-35.