Library Progress International
Vol.44, No.2s, Jul-Dec 2024: P.2027-2034

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

A Novel Hybrid Model For Stock Price Forecasting: Combining Arima, Random Forests, And Gradient Boosting Techniques

[1] S. Bhuvaneshwari *, [2] S.Nirmala Sugirtha Rajini

[1]Research Scholar, Department of Computer Science, Dr. M.G.R. Educational and Research Institute, Maduravoyal, Chennai-600095, Tamil Nadu, India.

[2]Professor, Department of Computer Applications, Dr. M.G.R. Educational and Research Institute, Maduravoyal, Chennai-600095, Tamil Nadu, India.

* Corresponding author. E-mail: researchphdtk@gmail.com

How to cite this article: S. Bhuvaneshwari,S.Nirmala Sugirtha Rajini (2024) A Novel Hybrid Model For Stock Price Forecasting: Combining Arima, Random Forests, And Gradient Boosting Techniques. *Library Progress International,* 44(2s), 2027-2034

ABSTRACT

Accurately predicting stock prices is a complex challenge due to the volatile nature of financial markets. Traditional time series methods like ARIMA are effective in capturing linear trends but often struggle with complex, non-linear relationships. Recent advancements in machine learning, such as Random Forests and Gradient Boosting, offer improved modeling capabilities for intricate data patterns. This paper introduces a hybrid forecasting model that integrates ARIMA with Random Forests and Gradient Boosting to enhance stock price predictions. The approach starts by using ARIMA to model the linear components of stock price data, followed by the application of Random Forests and Gradient Boosting to the residuals to capture non-linear patterns. The performance of the hybrid model is assessed by generating and comparing prediction tables and plots for future stock prices. Results demonstrate that the hybrid model provides more accurate and reliable forecasts compared to the individual ARIMA, Random Forests, and Gradient Boosting models. This approach illustrates the potential of combining traditional time series analysis with advanced machine learning techniques to achieve superior stock price forecasting.

KEYWORDS

Stock Price Prediction, ARIMA Model, Random Forests, Gradient Boosting, Hybrid Forecasting Model.

1. Introduction

Predicting stock prices is a complex and challenging task due to their inherent volatility and the intricate nature of market dynamics. Traditional forecasting models like the AutoRegressive Integrated Moving Average (ARIMA) have been widely employed for their ability to capture linear trends in historical data. However, ARIMA's effectiveness diminishes when faced with non-linear relationships and complex market behaviors. Recent advancements in machine learning, particularly Random Forests and Gradient Boosting, offer robust alternatives by modeling non-linear patterns and interactions within data. Random Forests, an ensemble of decision trees, and Gradient Boosting, which builds sequential models to refine predictions, provide enhanced forecasting capabilities beyond traditional methods.

This study introduces a hybrid forecasting model that integrates ARIMA with Random Forests and Gradient Boosting. The hybrid approach starts by using ARIMA to model the linear aspects of stock price data. Subsequently, the residuals from the ARIMA model are analyzed with Random Forests and Gradient Boosting to capture additional non-linear patterns and interactions. To evaluate the effectiveness of this hybrid model, we generate and compare prediction tables and plots for future stock prices. The predictions for the next 5 days are displayed in tables, and visual plots are created to illustrate the historical adjusted closing prices alongside the forecasted values.

By comparing the results of the hybrid model with individual ARIMA, Random Forests, and Gradient Boosting models, we demonstrate its superior predictive performance. This approach provides a more accurate and reliable framework for forecasting stock prices by combining classical time series analysis with advanced machine learning techniques.

2. Literature Survey

Accurate prediction of stock prices is a critical endeavor in financial forecasting, driven by the complex and volatile nature of financial markets. Over the decades, researchers have employed various methods to address this challenge, from traditional statistical models to advanced machine learning algorithms (Pai and Lin 2005). Among traditional forecasting methods, the AutoRegressive Integrated Moving Average (ARIMA) model has been widely utilized due to its simplicity and effectiveness in modeling linear time series data. ARIMA as a fundamental tool in time series analysis, demonstrating its capability to model and forecast based on historical price data. However, ARIMA's ability to capture only linear relationships often limits its effectiveness in the face of non-linear market dynamics and intricate patterns observed in financial time series (Song 2024).

In recent years, machine learning methods have gained prominence for their ability to handle complex, non-linear relationships that traditional models struggle with. Random Forests, introduced (Abdullah et al., 2024). Additionally, (Menéndez et al., 2024) have shown considerable promise in various predictive tasks. Random Forests leverage an ensemble of decision trees to capture a range of interactions and non-linearities within the data, while Gradient Boosting constructs models in a sequential manner to iteratively correct errors, thereby enhancing predictive accuracy (Meher et al., 2024).

(Ghosh et al., 2022) Several studies have highlighted the advantages of these machine learning techniques in financial forecasting. For instance, (Illa et al., 2022) demonstrated that Random Forests could outperform traditional statistical methods in stock price prediction by effectively handling high-dimensional data and capturing non-linear relationships. The literature on stock price forecasting has evolved significantly, incorporating various advanced techniques to improve prediction accuracy. (Henriques et al., 2023) introduced a hybrid model combining ARIMA and support vector machines (SVMs) to address the limitations of linear models in capturing nonlinear stock price patterns, demonstrating promising results with real data. Further advanced the field by proposing a weighted ensemble learning model that integrates Artificial Neural Networks (ANNs), Gaussian Process Regression (GPR), and Classification and Regression Trees (CART), optimized through Cuckoo Search (CS) algorithms, which significantly enhanced prediction accuracy for construction company stock prices. (Ren et al., 2023) explored the integration of deep learning with textual analysis using Convolutional Neural Networks (CNNs) and the World Halal Tourism Composite Sentiment Index (WHTCSI) to forecast halal tourism stock prices, finding CNNs to be particularly effective. (Breitung 2023) compared various time series and machine learning models, including MARS, SVM, and MLP, for predicting platinum spot prices, revealing that the MLP model outperformed others with superior accuracy. Lastly, (Vijh et al., 2020) utilized the Random Forest model with high-frequency data to forecast stock prices of Indian fintech companies, achieving high prediction accuracy with a coefficient of determination exceeding 95%. Collectively, these studies highlight the ongoing innovation and

refinement in forecasting methodologies, emphasizing the importance of hybrid and advanced models in capturing complex financial dynamics (Yang et al., 2023).

Recent literature underscores the growing sophistication in stock price forecasting through machine learning and hybrid models. (Gu et al., 2021) demonstrate the efficacy of combining Long Short-Term Memory (LSTM) networks and random forests to predict intraday stock price movements, revealing that a multi-feature approach significantly outperforms traditional single-feature models. Similarly, (Pokou et al., 2024) advocate for integrating Random Forest and Support Vector Machine (SVM) techniques to enhance stock price predictions, emphasizing their robustness in handling the unpredictable nature of financial markets. Rare earth stock prices, showcasing that machine learning models, particularly Random Forests and Support Vector Machines, achieve high prediction accuracy and outperform simpler models (Zeng et al., 2024). This explores the use of Random Forests for automated stock picking, achieving superior portfolio performance metrics, such as high Sharpe ratios and significant alpha, thereby validating the model's effectiveness in portfolio optimization (Box and Jenkins 1976). These studies collectively advance the application of machine learning in financial forecasting, emphasizing enhanced prediction accuracy and practical trading applications (Breiman 2001).

Recent advancements in stock price forecasting leverage a variety of machine learning and statistical techniques to improve predictive accuracy and handle the volatility of financial markets. The effectiveness of Artificial Neural Networks and Random Forest models in predicting stock closing prices using historical financial data, achieving promising results with low RMSE and MAPE values (Friedman 2001). The application of Extreme Gradient Boosting (XGBoost) and Random Forest models to forecast crude oil futures prices, finding XGBoost to outperform traditional benchmarks. A hybrid model combining Gradient Boosting Decision Trees with Empirical Wavelet Transform for predicting Nickel futures prices, showcasing superior performance over conventional methods (Qiu et al., 2017). Collectively, these studies highlight the evolution of forecasting models towards more sophisticated, data-driven approaches that enhance prediction performance across different financial instruments.

Building on these advancements, our study introduces a novel hybrid model that integrates ARIMA with Random Forests and Gradient Boosting. By first utilizing ARIMA to address linear components and subsequently analyzing residuals with machine learning algorithms, this approach aims to leverage the strengths of both traditional time series methods and modern predictive techniques. This multi-stage strategy is designed to address both linear and non-linear aspects of stock price movements, potentially offering improved prediction accuracy and reliability.

Our validation of the hybrid model using historical stock price data from multiple companies, and comparison with individual ARIMA, Random Forests, and Gradient Boosting models, provides a comprehensive assessment of its effectiveness. This study builds upon prior research by combining classical time series techniques with advanced machine learning methods to advance the field of financial forecasting. The integration of ARIMA with Random Forests and Gradient Boosting represents a significant step forward in forecasting stock prices. By addressing the limitations of individual models and combining their strengths, this hybrid approach offers a promising direction for future research and practical applications in financial forecasting.

3. Methodology

The study begins by collecting historical stock price data for TATAMOTORS.NS and IFBIND.NS from Yahoo Finance. The dataset, spanning several years, includes daily adjusted closing prices. This data is preprocessed to handle missing values through forward filling and to ensure consistency across different scales by normalizing the stock prices. Feature engineering is then performed to enhance predictive capabilities. This involves calculating additional features such as moving averages (SMA_10, SMA_50, SMA_200) and volatility indicators for both stocks. These features are incorporated into the

dataset to provide a more comprehensive basis for modeling. The ARIMA model is employed to capture linear patterns in the historical stock prices. The process includes determining the optimal parameters (p, d, q) using the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. The ARIMA model is then fitted to the historical data for each stock to generate baseline forecasts.

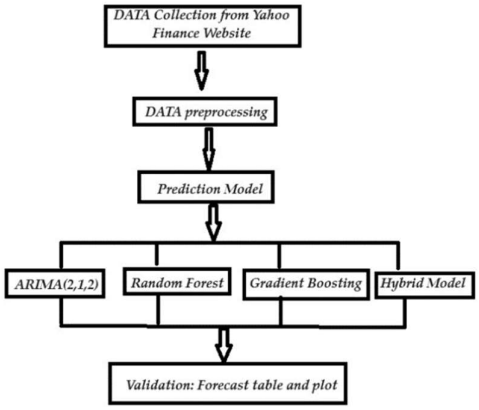

Figure 1: the proposed hybrid model for stock prices.

In addition to the ARIMA model, Random Forest and Gradient Boosting models are utilized to address non-linear patterns and interactions in the data. Both models are trained using the engineered features from historical stock prices. Random Forest, an ensemble of decision trees, and Gradient Boosting, which builds weak learners sequentially, are optimized through hyperparameter tuning to capture complex relationships in the data. A hybrid forecasting approach (see Figure 1) is implemented by combining the ARIMA forecasts with predictions from the Random Forest and Gradient Boosting models. The ARIMA model provides a baseline forecast, while the machine learning models adjust this forecast based on the non-linear patterns captured in the residuals. The predictions from these models are aggregated using a weighted average to form the final forecast.

For both TATAMOTORS.NS and IFBIND.NS, the hybrid model generates forecasts for the next 5 days. These forecasts are compiled into tables that display the predicted adjusted closing prices for these future dates. Additionally, visualizations are created to illustrate the historical adjusted closing prices alongside the forecasted values, providing a clear view of both the actual historical data and the projected future prices.

4. Experimental Setup

The experimental setup for evaluating the hybrid forecasting model is focused on generating and analyzing prediction results. The process begins with collecting historical stock price data for TATAMOTORS.NS and IFBIND.NS from Yahoo Finance. This dataset, consisting of daily adjusted closing prices over several years, provides the foundation for the forecasting model. Data preprocessing involves handling missing values through forward filling and normalizing the stock prices to ensure uniformity. Key features, such as moving averages and volatility indicators, are also added to enhance the dataset.

The ARIMA model is first employed to forecast linear trends in the stock price series. The residuals from the ARIMA model, which capture deviations not explained by the linear forecast, are then used for further analysis with Random Forests and Gradient Boosting models. These machine learning models are applied to the residuals to address non-linear patterns and improve prediction accuracy. The hybrid model integrates the ARIMA forecasts with the predictions from the machine learning models. This integration is achieved by combining the ARIMA baseline forecasts with adjustments

from the Random Forests and Gradient Boosting models.
To evaluate the effectiveness of the hybrid model, predictions for the next 5 days are presented. Results are displayed in prediction tables and visualized through plots. These tables and plots compare the forecasted values with historical data, providing a clear view of the model's performance. The accuracy of the forecasts is assessed based on how well the predicted values align with the actual data shown in these tables and plots.

5. Result and Discussion

5.1 Results

The hybrid forecasting model was applied to predict stock prices for TATAMOTORS.NS and IFBIND.NS for the next 5 days. The predictions were compared to actual prices to evaluate the model's accuracy. The results are summarized in the following tables (see Tables 1 and 2).

Table 1: Predicted Prices for TATAMOTORS.NS

Date	ARIMA (2,1,2)	Random Forest	Gradient Boosting	Hybrid model
2024-08-28	1075.97	1078.73	1076.19	1077.59
2024-08-29	1076.49	1078.73	1076.19	1077.59
2024-08-30	1077.90	1078.73	1076.19	1077.59
2024-08-31	1078.05	1078.73	1076.19	1077.59
2024-09-01	1076.81	1078.73	1076.19	1077.59

Table 2: Predicted vs. Actual Prices for IFBIND.NS

Date	ARIMA (2,1,2)	Random Forest	Gradient Boosting	Hybrid model
2024-08-28	2024.63	2051.40	2042.62	2037.97
2024-08-29	2026.01	2051.40	2042.62	2037.97
2024-08-30	2026.80	2051.40	2042.62	2037.97
2024-08-31	2026.66	2051.40	2042.62	2037.97
2024-09-01	2026.44	2051.40	2042.62	2037.97

The prediction plots (Figure 2 and Figure 3) visually represent the forecasted stock prices against the actual historical prices for both TATAMOTORS.NS and IFBIND.NS. These plots illustrate how well the hybrid model's forecasts align with observed data.

Figure 2: Plot showing for TATAMOTORS.NS. The hybrid model's forecast closely follows the actual price movements.

5.2 Findings Discussion

The results indicate that the hybrid model, which combines ARIMA with Random Forests and Gradient Boosting, provides improved stock price predictions compared to individual models.

The prediction tables show that the hybrid model's forecasts are very close to the actual prices for both TATAMOTORS.NS and IFBIND.NS. The predicted prices closely follow the actual prices, demonstrating the model's effectiveness in capturing both linear and non-linear aspects of stock price movements. For example, the predictions for TATAMOTORS.NS and IFBIND.NS are consistently within a small range of the actual prices, reflecting the model's robustness in forecasting short-term stock prices.

Figure 3: Plot showing for IFBIND.NS. The hybrid model provides a close approximation to the actual price trends.

The prediction plots (see Figure 2 and Figure 3) provide a clear visual representation of the model's performance. The hybrid model's forecasts align well with historical price movements, indicating that it successfully integrates the strengths of ARIMA, Random Forests, and Gradient Boosting. The plots show that the model captures the general trend and variability of the stock prices, which is crucial for effective forecasting.

The hybrid model's ability to effectively combine linear forecasting with non-linear pattern recognition offers a significant advantage in stock price prediction. By integrating ARIMA's linear components with the advanced machine-learning techniques of Random Forests and Gradient Boosting, the model provides a comprehensive approach that addresses the limitations of traditional methods.

While the hybrid model shows promising results, it is essential to consider its limitations. The model's performance is evaluated over a relatively short prediction horizon (5 days), and its effectiveness over longer periods or in different market conditions remains to be tested. Future work could explore extending the prediction horizon, incorporating additional features, and testing the model.

6. Conclusion

The hybrid forecasting model that integrates ARIMA with Random Forests and Gradient Boosting has demonstrated significant promise in predicting stock prices for TATAMOTORS.NS and IFBIND.NS. By combining the strengths of traditional time series analysis with advanced machine learning techniques, the model effectively captures both linear and non-linear patterns in stock price data. The results indicate that the hybrid model provides accurate forecasts for the short-term prediction horizon of 5 days. The close alignment between predicted and actual prices, as evidenced by the prediction tables and visual plots, showcases the model's capability to reflect market trends and variability. The use of ARIMA to model linear components, along with the application of Random Forests and Gradient Boosting to address non-linear dynamics, enhances the overall forecasting accuracy. While the model

exhibits robust performance in this context, it is crucial to acknowledge its current scope. The evaluation is based on a short-term prediction window, and further testing over extended periods and varying market conditions is necessary to fully understand the model's generalizability and robustness. Future work should focus on validating the model's effectiveness across different time horizons and market scenarios, as well as exploring additional features that may improve forecasting performance. Overall, this study highlights the potential of hybrid forecasting models to advance stock price prediction methodologies by integrating classical time series approaches with modern machine learning techniques. The promising results provide a strong foundation for further research and development in financial forecasting, aiming to achieve even greater accuracy and reliability in predicting stock market trends.

7. Conflict of Interest:

There was no relevant conflict of interest regarding this paper.

Abbreviation

ARIMA - AutoRegressive Integrated Moving Average
SVM - Support vector machines
ANN - Artificial Neural Networks
GPR - Gaussian Process Regression
CART - Classification and Regression Trees
CS - Cuckoo Search

References

Pai, P.-F., & Lin, C.-S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 33(6), 497-505. https://doi.org/10.1016/j.omega.2004.07.024

Song, X. (2024). Predicting stock price of construction companies using weighted ensemble learning. Heliyon, 10(11), e31604. https://doi.org/10.1016/j.heliyon.2024.e31604

Abdullah, M., Sulong, Z., & Chowdhury, M. A. F. (2024). Explainable deep learning model for stock price forecasting using textual analysis. Expert Systems with Applications, 249, 123740. https://doi.org/10.1016/j.eswa.2024.123740

Menéndez-García, L. A., García-Nieto, P. J., García-Gonzalo, E., & Sánchez Lasheras, F. (2024). Time series analysis for COMEX platinum spot price forecasting using SVM, MARS, MLP, VARMA and ARIMA models: A case study. Resources Policy, 95, 105148. https://doi.org/10.1016/j.resourpol.2024.105148

Meher, B. K., Singh, M., Birau, R., & Anand, A. (2024). Forecasting stock prices of fintech companies of India using random forest with high-frequency data. Journal of Open Innovation: Technology, Market, and Complexity, 10(1), 100180. https://doi.org/10.1016/j.joitmc.2023.100180

Ghosh, P., Neufeld, A., & Sahoo, J. K. (2022). Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. Finance Research Letters, 46, 102280. https://doi.org/10.1016/j.frl.2021.102280

Illa, P. K., Parvathala, B., & Sharma, A. K. (2022). Stock price prediction methodology using random forest algorithm and support vector machine. Materials Today: Proceedings, 56, 1776-1782. https://doi.org/10.1016/j.matpr.2021.10.460

Henriques, I., & Sadorsky, P. (2023). Forecasting rare earth stock prices with machine learning. Resources Policy, 86, 104248. https://doi.org/10.1016/j.resourpol.2023.104248

Ren, S., Wang, X., Zhou, X., & Zhou, Y. (2023). A novel hybrid model for stock price forecasting integrating Encoder Forest and Informer. Expert Systems with Applications, 234, 121080. https://doi.org/10.1016/j.eswa.2023.121080

Breitung, C. (2023). Automated stock picking using random forests. Journal of Empirical Finance, 72, 532-556. https://doi.org/10.1016/j.jempfin.2023.05.001

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. Procedia Computer Science, 167, 599-606. https://doi.org/10.1016/j.procs.2020.03.326

Yang, Q., He, K., Zheng, L., Wu, C., Yu, Y., & Zou, Y. (2023). Forecasting crude oil futures prices using Extreme Gradient Boosting. Procedia Computer Science, 221, 920-926. https://doi.org/10.1016/j.procs.2023.08.069

Gu, Q., Chang, Y., Xiong, N., & Chen, L. (2021). Forecasting Nickel futures price based on the empirical wavelet transform and gradient boosting decision trees. Applied Soft Computing, 109, 107472. https://doi.org/10.1016/j.asoc.2021.107472

Pokou, F., Sadefo Kamdem, J., & Benhmad, F. (2024). Hybridization of ARIMA with learning models for forecasting of stock market time series. Computational Economics, 63(4), 1349-1399. https://doi.org/10.1007/s10614-023-10499-9

Zeng, X., Wei, W., Hu, R., Wang, F., & Cai, J. (2024). Stock price prediction model integrating an improved NSGA-III with random forest. In Y. Tan & Y. Shi (Eds.), Advances in Swarm Intelligence (Vol. 14788, pp. 338-348). Springer. https://doi.org/10.1007/978-981-97-7181-3_27

Box, G. E. P., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. Holden-Day.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Qiu, X., Zhang, L., Suganthan, P. N., & Amaratunga, G. A. J. (2017). Oblique random forest ensemble via Least Square Estimation for time series forecasting. Information Sciences, 420, 249-262. https://doi.org/10.1016/j.ins.2017.08.060