Creative Economy
2025, VOL. 9, NO. 3, 106-111
DOI: 10.47297/wspceWSP2516-251921.20250903

Stock Price Prediction Based on ARIMA-LSTM Model: A Case Study of the Education Sector

Zihan Wang

Department of Financial Mathematics, Zhejiang University of Finance and Economics, Hangzhou, Zhejiang, 310000, China

ABSTRACT
In recent years, frequent reforms in education policies and the national emphasis on education-powered nation-building have led to significant fluctuations in the development of the education sector. Accurate prediction of financial time series has become a critical focus for investors. This study proposes a novel ARIMA-LSTM hybrid model to predict stock prices of the education sector index, Chuanzhi Education, and Guomai Media, and compares its performance with baseline ARIMA and LSTM models to evaluate predictive accuracy. Empirical results demonstrate that the ARIMA-LSTM model achieves superior accuracy in stock price forecasting compared to standalone ARIMA and LSTM models. By forecasting the development trends of the education sector, this study reveals that its future trajectory will remain relatively stable.

KEYWORDS
ARIMA model; LSTM model; Stock price prediction; Education sector

1 Introduction

With the rapid development of China's economy, the financial sector has experienced substantial growth, and the stock market, as a critical driver of economic progress, has played an increasingly vital role in fostering macroeconomic stability. It supports government efforts in both macro- and micro-economic regulation, thereby mitigating systemic financial risks. Consequently, stock price prediction has become a focal point for investors. However, forecasting stock prices remains highly challenging due to the complex interplay of influencing factors and the inherent characteristics of financial time series, including non-stationarity, high noise, and nonlinearity. Advances in technology have shifted predictive methodologies toward data-driven models and algorithms, particularly with the emergence of machine learning (ML) and deep learning (DL), which offer enhanced accuracy for investors. This study employs deep learning techniques to achieve scientifically robust stock price predictions.

The global education sector has emerged as a pivotal component of capital markets, driven by technological advancements and evolving economic priorities. As a cornerstone of societal development and economic growth, the education industry is undergoing transformative changes in its business models, notably through the rise of online education, vocational training, and educational technology (EdTech). Policy shifts, technological innovation, and shifting market demands significantly shape the sector's dynamics. Diverse educational needs—spanning K12 education, higher education, vocational training, and lifelong learning—continue to expand the market, with the global education sector projected to grow steadily in the coming years, particularly in emerging economies. However, education-related stocks exhibit heightened volatility due to regulatory interventions and technological disruptions. The 2025 national strategy to establish China as an "education powerhouse" has further intensified market fluctuations. Accurate prediction of education sector stock prices thus holds significant practical value for investors, enterprises, and policymakers.

The pursuit of achieving excess returns through stock investments has long been a central focus for investors, with accurate stock price prediction serving as a critical determinant. Traditional approaches to stock price forecasting primarily rely on fundamental and technical analyses, which evaluate a company's financial health, industry prospects, historical trading prices, and volumes. However, these methods are inherently subjective and dependent on the analyst's expertise, often yielding limited accuracy. Subsequent advancements in financial theory, including the Capital Asset Pricing Model (CAPM) by Sharpe, Markowitz's Mean-Variance Model, and Fama's Efficient Market Hypothesis, have laid the groundwork for more sophisticated predictive frameworks.

Early quantitative models for stock price prediction include the Autoregressive Integrated Moving Average (ARIMA) model (Chen et al., 2010 [1]; Shi et al., 2014 [2]; Wu et al., 2016 [3]), the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model (Yang et al., 2016 [4]; Huang et al., 2017[5]), wavelet prediction models (Lian et al., 2006 [6]; Bozic et al., 2018 [7]), multiple linear regression models, and factor analysis models (Zheng et al., 2015 [8]; Ning et al., 2024 [9]). Nevertheless, these models exhibit significant limitations, primarily applicable to short-term, stationary time series or linear relationships, thereby constraining their predictive efficacy.

The advent of internet technology and the proliferation of machine learning (ML) and deep learning (DL) methodologies have revolutionized predictive modeling in finance. These technologies enable automated trading strategies driven by big data analytics, representing a paradigm shift in modern investment practices. ML algorithms, particularly supervised learning techniques, excel in identifying complex patterns within nonlinear financial data by leveraging statistical inference and feature extraction. For instance, Zhang et al. (2016) [10]demonstrated that Support Vector Machine (SVM) models outperform traditional ARIMA-GARCH frameworks in denoising time series data, extracting trends, and enhancing prediction accuracy. Similarly, artificial neural networks (ANNs) have been widely adopted in financial forecasting, as evidenced by studies from Zhou et al. (2019) [11] and Zhang et al. (2024) [12].

Despite these advancements, single-model approaches often struggle to capture the full complexity of financial markets. To address this, researchers have increasingly turned to hybrid models that integrate multiple algorithms. Zhou (2021) [13] proposed a random forest-based multi-factor model for predicting CSI 500 index returns, exemplifying the synergistic integration of artificial intelligence and traditional financial analysis. Meanwhile, DL architectures such as Recurrent Neural Networks (RNNs) have shown promise in modeling temporal dependencies, though their practical utility was initially limited by gradient vanishing/exploding issues. The introduction of Long Short-Term Memory (LSTM) networks by Hu (2021) [14] mitigated these challenges, significantly improving prediction efficiency. Zhang (2024) [15] further validated LSTM's superiority over RNN in forecasting non-stationary stock prices, such as those of Tata Global Beverages.

Recent innovations in hybrid DL models include the integration of Convolutional Neural Networks (CNNs) with LSTMs to capture both spatial and temporal features. Geng et al. (2021) [16] demonstrated that CNN-LSTM models outperform standalone architectures in financial time series forecasting. Additionally, Zhang (2003) [17] pioneered the combination of ARIMA with ANNs, showcasing enhanced predictive stability. Advanced hybrid frameworks, such as the LSTM-Markov chain model proposed by Yuan et al. (2021) [18] and the RF-MIP-LSTM model developed by Zhang et al. (2024) [19], have consistently demonstrated improved accuracy across diverse datasets, including the stock performance of Gree Electric Appliances.

In summary, DL algorithms exhibit superior adaptability and accuracy in stock price forecasting compared to traditional models, with hybrid frameworks achieving state-of-the-art performance. However, while ARIMA excels in modeling linear trends and LSTM in capturing nonlinear dynamics, their standalone applications face limitations. This study proposes a novel ARIMA-LSTM hybrid model that decomposes time series into linear and nonlinear components, aiming to enhance prediction accuracy for education sector stocks. This approach not only addresses the limitations of single-model frameworks but also provides actionable insights for stakeholders navigating the volatile education market.

2 Model construction

2.1 ARIMA model

The ARIMA model, typically denoted as ARIMA(p, d, q), integrates autoregressive (AR) and moving average (MA) components. Here, p represents the lag order of the AR model, d denotes the differencing order required to achieve stationarity, and q signifies the lag order of the MA model. The general expression for ARIMA is:

$$\Phi(B)\nabla^d y_t = \Theta(B)\epsilon_t$$

where:

y_t: Time series value at time t,

∇^d: Differencing operator of order d,

$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$: Autoregressive polynomial of order p,

$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$: Moving average polynomial of order q,

B: Lag operator,

ϵ_t: White noise error term.

ARIMA Modeling Procedure:

a. Stationarity TestPerform stationarity checks (e.g., Augmented Dickey-Fuller (ADF) test on the original data. If non-stationary, apply differencing until stationarity is achieved, determining the optimal differencing order d.

b. Order DeterminationIdentify preliminary p and q values via autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Finalize orders by minimizing Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

c. Parameter EstimationEstimate coefficients and ϕusing θmaximum likelihood estimation (MLE) or moment methods, followed by statistical validation.

d. Residual DiagnosticsValidate residuals for white noise properties (e.g., Ljung-Box test). If residuals are non-random, re-specify the model.

2.2 LSTM model

Long Short-Term Memory (LSTM), proposed by Hochreiter and Schmidhuber (1997), addresses the vanishing gradient problem in recurrent neural networks (RNNs). An LSTM unit comprises three gating mechanisms—forget gate, input gate, and output gate—along with a memory cell.

a.Forget GateThe forget gate determines which information to discard from the memory cell. It processes the current input x_t and previous hidden state $h_{t-1,}$through a sigmoid function, generating a vector$f_t \in [0,1]$, which scales the memory cell state C_{t-1}:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1,} x_t\right] + b_f\right)$$

b.Input GateThe input gate updates the memory cell by selecting new information to store. It employs sigmoid and tanh activations to generate an input modulation vector $\widetilde{C}_t$:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1,} x_t\right] + b_i\right)$$

$$\widetilde{C}_t = \tanh\left(W_c \cdot \left[h_{t-1,} x_t\right] + b_c\right)$$

c.Memory Cell UpdateThe memory cell state C_t is updated by combining filtered historical information and new inputs:

$$C_t = f_t . C_{t-1} + i_t . \widetilde{C}_t$$

d.Output GateThe output gate regulates the hidden stateh_t, which is derived from the updated memory cell state:

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1,} x_t\right] + b_o\right)$$

$$h_t = o_t . \tanh\left(C_t\right)$$

2.3 ARIMA-LSTM hybrid model

The ARIMA-LSTM hybrid model synergizes ARIMA's linear modeling capability with LSTM's nonlinear feature extraction. It decomposes the original data into linear components and residuals, where ARIMA captures linear trends, and LSTM models nonlinear residuals.

Implementation Steps:

a.DecompositionDecompose the original time series y_t into linear predictions $\widetilde{ARIMA}$and residuals ε_t:

$$\varepsilon_t = y_t - \widetilde{ARIMA}$$

b.Hybrid Forecasting

ARIMA predicts the linear component $\widetilde{ARIMA}$.

LSTM models the nonlinear residualsε_t , generating $\widetilde{LSTM}$.

c.Final PredictionCombine outputs to obtain the hybrid forecast:

$$\widetilde{y}_t = \widetilde{ARIMA} + \widetilde{LSTM}$$

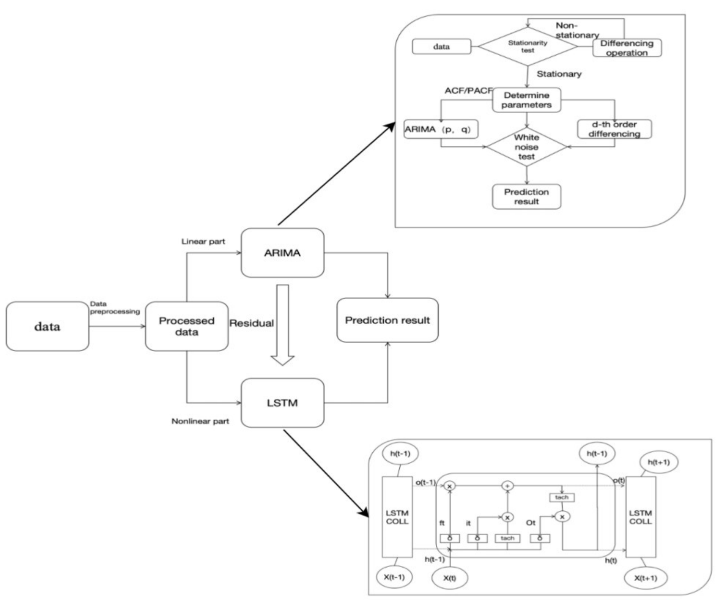

Figure 1

2.4 Model evaluation metrics

To objectively assess model performance, this study employs three evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Their mathematical formulations are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right|$$

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2\right)^{1/2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \widehat{y_i}\right|}{y_i} \times 100\%$$

3 Empirical analysis

3.1 Data selection and preprocessing

To ensure compliance with the characteristics of financial time series, this study utilizes closing prices, trading volumes, returns, turnover rates, price-to-earnings (P/E) ratios, price-to-book (P/B) ratios, relative strength index (RSI), and moving average convergence divergence (MACD) data from three listed companies in China's education sector. Specifically, historical data of Guomai Technology, DouShen Education, and Zhonggong Education from the first trading day of 2005 to the last trading day of 2024 were analyzed. The dataset for each company exceeds 3,000 observations, with minor discrepancies in sample sizes due to varying listing dates. All data were sourced from the China Center for Economic Research (CCER) Economic and Financial Database to ensure reliability and consistency.

3.2 Model parameter configuration

The forecasting models were implemented in Python 3.12 and TensorFlow 2. x, using Jupyter Notebook as the development environment.

ARIMA Parameter Selection:The optimal parameters (p, d, q) for the ARIMA model were determined by minimizing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

LSTM Architecture:

a.Input Features:

For the standalone LSTM model: Preprocessed raw data (normalized time series).

For the hybrid ARIMA-LSTM model: Residuals from the ARIMA model combined with selected financial indicators.

b.Network Structure:

Bidirectional LSTM with two hidden layers (80 cells per layer).

A Dropout layer (rate = 0.3) was appended after each hidden layer to mitigate overfitting.

c.Training Configuration:

Optimizer: Adam with default hyperparameters.

Batch size: 32.

Training epochs: 500.

3.3 Empirical results analysis

3.3.1 Closing price prediction analysis for Zhonggong education

This study first evaluates the predictive performance of different models using the closing price data of Zhonggong Education, covering the period from August 10, 2011, to December 31, 2024 (3,255 data points). The forecasting accuracy is quantified via three metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Comparative results across models are summarized inTable 1 .

According to the results in Table 1, for the stock price prediction of Zhonggong Education, the ARIMA model achieved an MAE of 0.18116, while the LSTM and ARIMA-LSTM models yielded higher MAE values of 0.59546 and 0.18181,

Table 1

model	MAE	RMSE	MAPE(%)
ARIMA	0.18116	0.13314	3.65
LSTM	0.59546	0.43122	16.63
ARIMA-LSTM	0.18181	0.13140	3.58

respectively. In contrast, the RMSE values for the ARIMA and LSTM models were 0.13314 and 0.43122, whereas the ARIMA-LSTM hybrid model demonstrated superior performance with an RMSE of 0.13140. Similarly, the ARIMA-LSTM model outperformed both baseline models in terms of MAPE, achieving a value of 3.58%, compared to the higher MAPE values of the standalone ARIMA and LSTM models.

3.3.2 Closing price prediction analysis for Guomai technology

This study further evaluates the predictive performance of the models using historical closing price data of Guomai Technology, spanning from December 15, 2006, to December 31, 2024 (4,386 observations). The forecasting accuracy of ARIMA, LSTM, and ARIMA-LSTM models is quantifiedthrough MAE, RMSE, and MAPE metrics, as detailed in Table 2.

Table 2

model	MAE	RMSE	MAPE(%)
ARIMA	0.28309	0.18159	2.57
LSTM	0.34778	0.22154	3.16
ARIMA-LSTM	0.28392	0.18115	2.56

According to Table 2, the ARIMA model achieved an MAE of 0.28309 for Guomai Technology, while the LSTM and ARIMA-LSTM models yielded higher MAE values of 0.34778 and 0.28392, respectively. For RMSE, the ARIMA and LSTM models produced values of 0.18159 and 0.22154, whereas the ARIMA-LSTM hybrid model demonstrated marginally lower RMSE (0.18115), indicating slight improvement over ARIMA. Similarly, the ARIMA-LSTM model outperformed both baseline models in MAPE, achieving a value of 2.56%, compared to the higher MAPE values of the standalone ARIMA and LSTM models.

3.3.3 Closing price prediction analysis for Doushen education

This study further examines the forecasting performance of the models using historical closing price data of Doushen Education, covering the period from October 30, 2009, to December 31, 2024 (3, 686 observations). The predictive accuracy of ARIMA, LSTM, and ARIMA-LSTM models is evaluated through MAE, RMSE, and MAPE metrics, as summarized in Table 3.

Table 3

model	MAE	RMSE	MAPE(%)
ARIMA	0.39883	0.18626	4.49
LSTM	0.96157	0.71283	22.72
ARIMA-LSTM	0.40020	0.18590	4.47

According to Table 3, the ARIMA model achieved an MAE of 0.39883 for Doushen Education, while the LSTM and ARIMA-LSTM models exhibited higher MAE values of 0.96157 and 0.40020, respectively. In terms of RMSE, the ARIMA and LSTM models yielded values of 0.18626 and 0.71283, whereas the ARIMA-LSTM hybrid model demonstrated a marginally lower RMSE of 0.18590, indicating slight improvement over the standalone ARIMA model. Similarly, the ARIMA-LSTM model outperformed both baseline models in MAPE, achieving a value of 4.47%, compared to the higher MAPE values of the ARIMA and LSTM models.

4 Conclusions and limitations

This study proposes a hybrid ARIMA-LSTM model that integrates linear and nonlinear components, incorporating key financial indicators—including trading volume, returns, turnover rate, price-to-earnings (P/E) ratio, price-to-book (P/B) ratio, relative strength index (RSI), and moving average convergence divergence (MACD)—to forecast stock closing prices. Through comparative analyses across three models (ARIMA, LSTM, and ARIMA-LSTM) and multiple stocks, the ARIMA-LSTM hybrid model demonstrates superior performance in predicting closing prices. Specifically, it outperforms standalone ARIMA and LSTM models in terms of RMSE and MAPE. While the hybrid model achieves lower MAE values than the LSTM model, its MAE performance relative to ARIMA is comparable. Overall, the ARIMA-LSTM framework enhances predictive accuracy for stock prices.

However, the proposed hybrid model has limitations. First, limited data availability—due to varying listing dates of the selected stocks—constrains the training efficacy of the LSTM component, potentially reducing prediction precision. Second, the current integration strategy, which simply combines ARIMA predictions with LSTM-modeled residuals through additive aggregation, offers only marginal improvements. Future studies could explore alternative integration strategies (e. g., weighted combinations, dynamic gate mechanisms) to better capture complex interactions between linear and nonlinear patterns, thereby improving adaptability to diverse financial time series.

About the Author

* Corresponding Author : Zihan Wang, 1044052460@qq.com.

References

[1] Chen L., Huang Z. S. Futures price analysis and forecasting based on ARIMA model. Journal of Fuzhou University (Philosophy and Social Sciences Edition),2010,24(3):32-37.

[2] Shi H.Y., You Z.J., Chen Z.J. Analysis and prediction of Shanghai Composite Index based on wavelet-ARIMA model. Mathematics in Practice and Theory,2014,44(23):66-72.

[3] Wu Y.X., Wen X. Short-term stock price prediction based on ARIMA model. Statistics and Decision,2016(23):83-86.

[4] Yang Q., Cao X.B. Stock price analysis and forecasting based on ARMA-GARCH model. Mathematics in Practice and Theory,2016,46(6):80-86.

[5] Huang X.Z., Song W., Liu Z.Y. Exchange rate volatility forecasting based on Markov-switching GARCH model. Statistics and Decision,2017(6): 84-87.

[6] Lian X.Y., Wu L.F. Stock price prediction based on wavelet neural network. Market Modernization,2006(29):331-332.

[7] Bozic J., Babic D. Financial time series forecasting using hybrid wavelet-neural model. International Arab Journal of Information Technology, 2018,15(1):50-57.

[8] Zheng R., Zhou L. Stock price prediction combining factor analysis and neural network. Journal of Inner Mongolia Normal University (Natural Science Edition),2015,44(4):452-456.

[9] Ning H., Chen Y. N., Yang L. J., et al. Stock portfolio optimization based on factor analysis and second-order memetic differential evolution algorithm. Memetic Computing,2024,16(1):29-44.

[10] Zhang G. S., Zhang X. D. SVM-GARCH stock price prediction model based on nearest neighbor mutual information. Chinese Journal of Management Science,2016,24(9):11-20.

[11] Zhou X.B., Chen Z., Wang J.Y. RMB exchange rate forecasting based on hybrid artificial neural network: A comparison with ARMA, ARCH, and GARCH. International Economics and Trade Research,2019,35(9):35-49.

[12] Zhang Han, Wang Weiguo. Neural Network Ensemble Model for Financial Time Series Prediction [J/OL]. Journal of Beijing University of Posts and Telecommunications,1-6[2025-02-07].

[13] Zhou L.Research on stock multi-factor investment based on random forest model. Financial Theory and Practice,2021(7):97-103.

[14] Hu Y.W. Stock prediction based on optimized LSTM model. Computer Science,2021,48(S1):151-157.

[15] Zhang X.F. Application of LSTM model in stock price prediction based on deep learning. Technology and Innovation,2024(23):23-26.

[16] Geng J.J., Liu Y.M., Li Y., et al. Stock index prediction model based on CNN-LSTM. Statistics and Decision,2021,37(5):134-138.

[17] Zhang G.P. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing,2003,50:159-175.

[18] Yuan G.L., Han H., Tang H.Q., et al. Stock price prediction based on LSTM neural network-Markov chain composite model. Guangxi Economy, 2022,40(5):88-93.

[19] Zhang Y., Li L. RF-MIP-LSTM stock price prediction model. Computer Engineering and Applications,2024,60(17):272-281.