ISSN 2616-5775 Vol. 7, Issue 11: 127-134, DOI: 10.25236/AJCIS.2024.071117

A Novel Residual Correction Approach Based on a Hybrid GARCH and XGBoost Model

Jiahao Wang[1,a,*], Jihong Zhang[2,b], Yuxin Cai[1,c]

[1]Department of Mathematics and Statistics, Northwest Normal University, Lanzhou, China
[2]Department of Information Technology, Lanzhou City University, Lanzhou, China
[a]2022221989@nwnu.edu.cn, [b]cyxcmz@163.com, [c]zhjhzhangjihong@163.com
[*]Corresponding author

Abstract: *The stock market's inherent uncertainty and volatility make forecasting stock prices a highly challenging task in finance. Financial time series modeling plays a crucial role in addressing this issue. This study proposes a hybrid model combining ARIMA, GARCH, and XGBoost to improve the accuracy of stock opening price predictions. The hybrid approach adjusts ARIMA model residuals using GARCH and XGBoost, leveraging their complementary strengths. Historical opening price data from Hong Kong stocks were employed to train and validate the model. Comparative analyses with traditional single models and other hybrid models reveal that the proposed model demonstrates superior accuracy and robustness. These results highlight its potential to provide reliable decision-making support for investors in the stock market.*

Keywords: *Stock Price Forecasting, ARIMA-GARCH model, XGBoost model, Residual Correction*

1. Introduction

As the global financial market continues to evolve, the stock market, one of the most complex and dynamic financial arenas, has consistently attracted the attention of investors and researchers. Stock price fluctuations are influenced by numerous factors, including macroeconomic indicators, market sentiment, company performance, and other variables, making stock price prediction a challenging task. In this era of information proliferation, investors increasingly demand precise and timely stock price forecasts to enhance their strategies and mitigate risks. Forecasting stock prices remains a significant challenge in finance due to market nonlinearity and uncertainty. Traditional methods, which rely heavily on statistical modeling and technical analysis, often perform poorly in predicting future trends. The emergence of deep learning technologies, especially the widespread adoption of artificial neural networks, has created new opportunities to advance this field. Wu Yuxia et al.[1] developed an ARIMA model to forecast fluctuations in the GEM market, achieving superior short-term predictive capabilities. Chen Yushao et al.[2] integrated the XGBoost algorithm with Pearson optimization to address data complexity and long training times associated with traditional neural networks, achieving efficient closing price forecasts. Peng Yan et al.[3] applied the LSTM methodology, resulting in low computational complexity and high predictive precision. Yuan et al.[4]formulated a PCA-BP-based forecasting model, which was validated with out-of-sample data and demonstrated enhanced predictive accuracy after PCA implementation.

However, these studies did not consider residual correction after model prediction. In this study, we use the traditional ARIMA method to forecast stock price movements and apply the GARCH and XGBoost methods to correct residuals from the ARIMA model. Consequently, this paper proposes a composite model for predicting stock opening prices based on ARIMA-GARCH-XGBoost.

2. Proposed Methods

2.1. ARIMA Model

The ARIMA (Autoregressive Integrated Moving Average) model is a statistical tool utilized for time series analysis and forecasting.[5] It comprises autoregressive (AR), moving average (MA), and differencing components. The fundamental concept of the ARIMA model involves treating the time-dependent data series as a collection of random variables dependent on time 't', capturing trends and seasonality through observation of autocorrelation and moving averages, thereby enabling future value prediction. ARIMA models are primarily suitable for analyzing smooth time series data; non-smooth series require transformation via differencing to facilitate subsequent operations.[6] The key parameters of an ARIMA model are denoted as p, d, q, which can be leveraged to forecast future data values based on observations of autocorrelation and moving averages. In the MA (Moving Average) component, the model employs a linear combination of past prediction errors to forecast future values. The parameter q represents the number of past prediction errors utilized in the model, i.e., the moving average order.

The ARIMA model is derived by integrating the aforementioned components. An ARIMA(p,d,q) model signifies that the original sequence undergoes d-order differencing for smoothing, and the resulting differenced sequence can be represented as a linear combination of the previous p observations and q residuals.[7]

$$ARIMA(p,d,q): Y_t = \mu + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - - \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{1}$$

The general procedure for constructing an ARIMA model involves initially assessing the smoothness of the original series and applying differencing if it is non-smooth. Subsequently, the appropriate order of the model (i.e., selecting potential p and q values) is determined based on autocorrelation and partial autocorrelation plots. The model is then fitted, and parameters are estimated using specific evaluation criteria such as AIC and BIC values to identify the optimal model. Following this, rigorous testing and diagnosis of the model are conducted before utilizing it for forecasting purposes.

2.2. GARCH Model

The GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model is an econometric framework commonly employed in quantitative research to account for and predict heteroskedasticity (conditional heteroskedasticity) in financial and economic time series data.[8] This model is grounded in two fundamental principles.

Firstly, the GARCH model incorporates the autoregressive (AR) concept to capture the mean component of time series data. The autoregressive model assumes a correlation between current and previous observations, enabling prediction of future mean levels based on autocorrelation in historical data. Secondly, by introducing Conditional Heteroskedasticity, the GARCH model aims to capture the volatility pattern of time series data.[9] Conditional heteroskedasticity signifies that data variance fluctuates over time, making it a function of time. The objective of the GARCH model is to enhance accuracy in risk assessment and volatility forecasting through modeling and predicting conditional heteroskedasticity.

2.3. XGBoost Model

XGBoost (eXtreme Gradient Boosting) is a robust machine learning model rooted in gradient boosting trees, renowned for its prowess in predictive modeling and feature selection. XGBoost models are engineered to enhance predictive performance through iterative training of multiple weak learners (typically decision trees), which are then amalgamated into a potent single learner used for delivering final prediction results, thereby elevating prediction accuracy. XGBoost boasts several pivotal features.

Firstly, XGBoost is built upon gradient boosting trees, where the basic weak learner consists of a series of decision trees trained iteratively to compensate for prediction errors and improve data fitting through step-by-step iteration.[10] Secondly, it optimizes the objective function by minimizing the loss function using gradient boosting and adding a regularization term to control model complexity and mitigate overfitting risks. Thirdly, XGBoost features capability for evaluating feature importance, facilitating effective feature selection to enhance model efficiency in predictive tasks. Lastly, it supports parallelization by enabling training with multiple threads and massively parallel computation in distributed environments, thereby accelerating model training significantly. The expression for the objective function of XGBoost is depicted in equation (2).

$$\mathcal{L}^{(t)} = \sum_i l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \sum_k \Omega\left(f_k\right) \tag{2}$$

where $\sum_k \Omega(f_k)$ is the regularization term, with the specific expression $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ and γ, λ are the regularization parameters. The expression of the optimal objective function, derived from leaf node traversal and utilizing second order Taylor expansion, is presented in equation (3).

$$\mathcal{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{{G_j}^2}{H_j + \lambda} + \gamma T \tag{3}$$

The formul $G_j = \sum_{i \in I_j} g_i$, $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$, $H_j = \sum_{i \in I_j} h_i$ and $h_i = \partial^2{}_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$.

2.4. ARIMA-GARCH-XGBoost Model

The ARIMA-GARCH-XGBoost model is a composite approach that integrates the ARIMA, GARCH, and XGBoost models. The GARCH model is used to characterize the ARIMA model when the residual term of the ARIMA model exhibits substantial conditional heteroskedasticity.[11]Subsequently, the residuals from the ARIMA model are forecasted using the XGBoost model, incorporating additional external features to capture more complex information inherent in the time-series data. Finally, a linear regression is applied to combine the residual predictions from both the GARCH and XGBoost models, thereby correcting the residuals of the ARIMA model.

3. Stocking Opening Price Modeling

The obtained sequence of stock opening prices is modeled as follows: Step 1: Generate a time series plot of the original stock opening price sequence to observe its trend and assess its smoothness. Alternatively, conduct an Augmented Dickey-Fuller (ADF) test to determine the smoothness of the original sequence. If the original sequence is found not to be smooth, apply differencing to achieve smoothness. Subsequently, perform a white noise test on the differenced smooth sequence and evaluate its status as a white noise sequence based on the p-value. A non-white noise result from this test indicates research significance for the sequence. Step 2: Generate autocorrelation (ACF) and partial autocorrelation (PACF) plots of the differenced series to initially ascertain the order of the ARIMA model. Subsequently, employ the AIC criterion to compare the AIC values of ARIMA models with different orders, determine the optimal model order, and derive the optimal ARIMA model. Step 3: Employ the ARIMA model to obtain the predicted values and compute the residuals between the predicted and actual values to derive the residual series. Step 4: Heteroskedasticity test for the residual series, i.e., an ARCH effect test. If the residual series exhibits heteroskedasticity, a GARCH model should be constructed to correct the residuals for prediction. Step 5: An XGBoost model is constructed for the residual series to capture the nonlinear relationship within the residuals and refine them for prediction. Step 6: Develop a linear regression model to integrate the predictive outcomes of GARCH and XGBoost models on the residual series, obtain the optimal predictive results, and achieve the ultimate refinement of residuals. Step 7: The ARIMA model predictions in step 3 are combined with the residual predictions in step 6 to derive the stock opening price predictions of the ARIMIA-GARCH-XGBOOST composite model.

The model's flowchart is depicted in Figure 1.

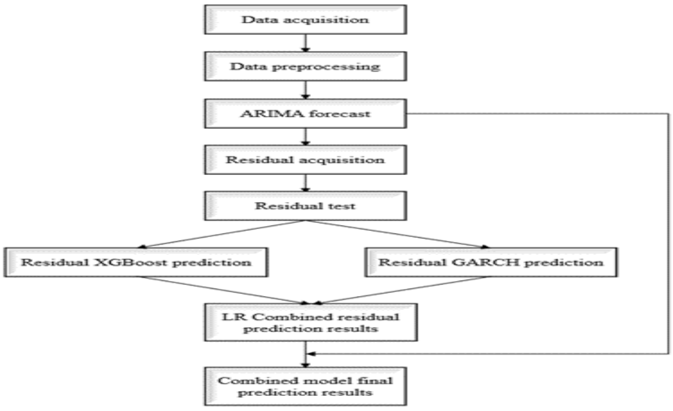
Data acquisition
Data preprocessing
ARIMA forecast
Residual acquisition
Residual test
Residual XGBoost prediction
Residual GARCH prediction
LR Combined residual prediction results
Combined model final prediction results

Figure 1: Flowchart of the ARIMA-GARCH-XGBoost model

4. Empirical Studies

4.1. Data Sources

This study constructs a model to select stocks for the Hong Kong stock Shangtang-W (HK00020). Shangtang Group is recognized as Asia's largest artificial intelligence software company, holding significant market influence. The dataset consists of daily opening price data for HK00020 stock in 2022, excluding weekends, totaling 246 entries with no missing values or outliers. The data was sourced from the publicly available dataset on the Kaggle competition platform. Data analysis and modeling in this study were conducted within a Python 3.7 environment using the scikit-learn library.

4.2. Stability Check

Prior to ARIMA modeling, a test for smoothness should be conducted on the original series to assess whether the time series data exhibits characteristics of a smooth distribution.[12] The smoothness of the sequence can be determined based on both the original sequence graph and ADF test; if the original sequence lacks smoothness, differencing is necessary to achieve a smoother sequence. Figure 2 illustrates the time series plot of the original opening price sequence.

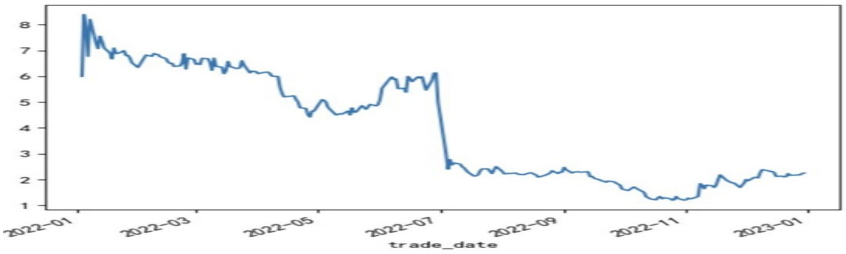

Figure 2: Timing chart of the original opening price

Based on the figure, it is evident that the original series exhibits a declining trend and lacks smoothness. Furthermore, the Augmented Dickey-Fuller (ADF) test results indicate a p-value of 0.68 for the original sequence, exceeding 0.05. Consequently, we reject the null hypothesis and conclude that the original sequence is non-stationary, necessitating differentiation. Subsequently, the ADF test yields a p-value of 0.0136 for the differenced series, supporting our initial hypothesis that this differentiated series demonstrates smoothness and can be effectively modeled using ARIMA.

4.3. Model Selection and Parameter Estimation

The differenced series represents a smooth non-white noise sequence, and the optimal order of the ARIMA model must be determined during the model selection phase.[13]By examining the autocorrelation and partial autocorrelation diagrams of the differenced series, as well as applying the principle of minimum information content AIC, we determine that the optimal parameters p,d,q are 4,1,3; thus, indicating that the ARIMA(4,1,3) model is best suited for predicting opening price trends. Figure 3 illustrates the fitting effect of ARIMA(4,1,3) on the differenced series.

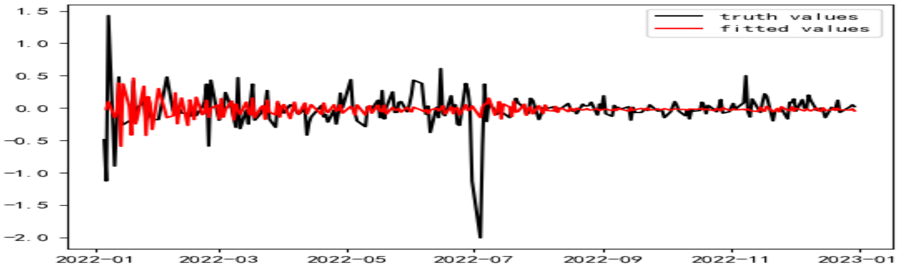

Figure 3: Effect of sequence fitting after ARIMA differencing

Based on the fitting effect graph of ARIMA(4,1,3), it is evident that the residuals exhibit a certain level of clustering, which exerts an influence on the accuracy of prediction results. Therefore, it becomes imperative to rectify the residuals to enhance the predictive model's precision.

4.4. XGBoost Correction of Residuals

After obtaining the ARIMA (4,1,3) prediction results, the model's residuals can be derived by comparing them with the true values. Subsequently, the XGBOOST model is employed to rectify these residuals.[14] During this process, consideration is given to the correlation between other features and the residuals, as depicted in Figure 4 which illustrates a correlation heat map of each variable.

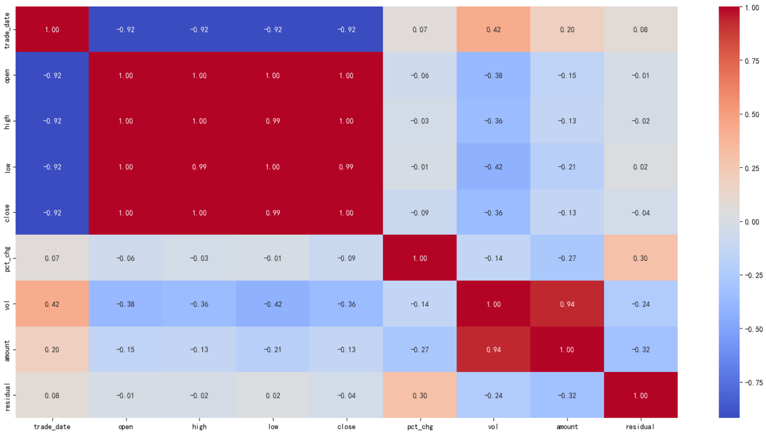

Figure 4: Heat map of correlation of variables

From the figure, it can be seen that there is a correlation between the residuals and pct_chg,vol and AMOUNT, Therefore, these three variables pct_chg,vol and AMOUNT are selected as features while utilizing the residuals as labels to construct an XGBoost model for predicting the residuals. In this model, 70% of the data is allocated to training and testing is performed on the remaining 30%. Following parameter tuning, Table 1 displays the optimal hyperparameters for the XGBoost model.

Table 1: Optimal hyperparameters for the XGBoost model

parameters	optimum value
n_estimators	40
max_depth	3
min_child_weight	1
Gamma	0.2
colsample_bytree	0.6
Subsample	0.7
reg_alpha	0.01
learning_rate	0.1

4.5. GARCH Correction of Residuals

In addition to employing the XGBoost model for residual correction, an ARCH effect test was conducted on the previous ARIMA(4,1,3) model. The test revealed heteroskedasticity in the residual series, indicating that a GARCH model could be utilized to enhance the ARIMA model and rectify residuals for predictive purposes.

The mean equation utilizes ARIMA(4,1,3), and the optimal order of the GARCH model is determined based on the AIC value. When p=q=1, the AIC value reaches its minimum. Consequently, we construct an ARIMA(4,1,3)-GARCH(1,1) model and present the correlation coefficients in Table 2.

Table 2: Sexual correlation coefficients for the GARCH model

index ratio		coefficient estimate	standard deviation	P-value
mean value	φ	-0.2077	0.09224	0.02431
variance	ω	0.00619	0.0048	0.197
	α_1	0.6238	0.328	0.04682
	β_1	0.3762	0.233	0.107

In this study, cross-validation and grid search techniques were employed to optimize the hyperparameters of Random Forest, XGBoost, and DNN models. By comparing the predictive performance on the test set, we identified the optimal hyperparameters for each high-quality model. The specific optimal hyperparameters for each model are presented in Table 2.

From the table 2, it is evident that the coefficient of the ARCH term contracts, indicating short-term autocorrelation in the series. The fluctuations in the current period's stock opening price can significantly impact short-term fluctuations. The combined coefficientsα1andβ1 in the variance equation demonstrate strong persistence of external factors and past stock price volatility on future stock prices, suggesting stability in the equation. Thus, the constructed ARIMA(4,1,3)-GARCH(1,1) model is valid for forecasting stock opening prices.

4.6. GARCH Correction of Residuals

In order to evaluate and compare the performance of the prediction models, this study employs three widely used evaluation metrics for regression problems: mean absolute percentage error (MAPE), mean square error (MSE), and coefficient of determination (R^2). It is assumed that the true temperature value for the samples in the test set is denoted as y_i, while the predicted value obtained from the model is represented as $\hat{y}_i$. The specific formulas for calculating MAPE, MSE, and R^2 are as follows:

$$\boldsymbol{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{4}$$

$$\boldsymbol{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \tag{5}$$

$$\boldsymbol{R}^2(\boldsymbol{y},\hat{\boldsymbol{y}}) = 1-\frac{\sum_{i=0}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{N}(y_i-\bar{y})^2} \tag{6}$$

The evaluation metrics for each of the ARIMA, AIRIMA-GARCH, ARIMA-XGBOOST, and ARIMA-GARCH-XGBoost models based on the test set data for model comparison are presented in Table 3.

Table 3: Comparison of evaluation indicators of the three models

model evaluation indicators	MAPE	MSE	R^2
ARIMA	0.05341	0.015869	0.88982
ARIMA-GARCH	0.00961	0.000599	0.99584
ARIMA-XGBoost	0.00981	0.000989	0.99582
ARIMA-GARCH-XGBoost	0.00754	0.000284	0.99799

The comparison reveals that the overall predictive performance of the model is superior, with the combined ARIMA-GARCH-XGBoost model demonstrating the most effective prediction among all models.[15]The three indices exhibit significant improvement compared to the ARIMA model, indicating that the residual component of the ARIMA model can be effectively corrected by the GARCH-XGBoost model to capture the fluctuation characteristics and enhance predictive accuracy. To visually demonstrate each model's fitting, a comparative graph is presented in this paper depicting real stock opening price data against predicted data for each model. This graphical representation clearly illustrates that the predicted curve of ARIMA-GARCH-XGBoost closely aligns with actual trends when compared to other models, as depicted in Figure 5 and Figure 6.

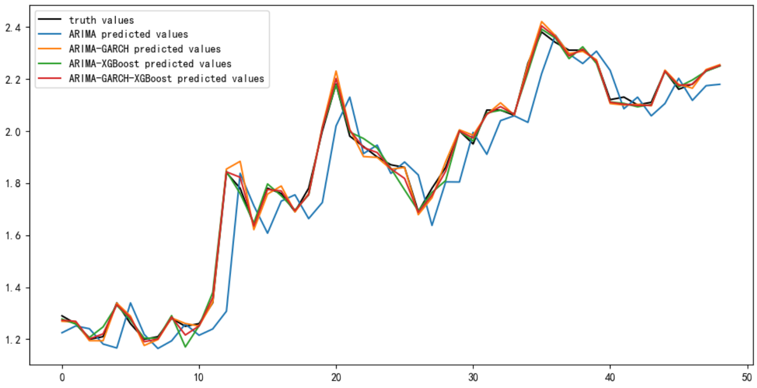

Figure 5: Comparison of the predicted results of each model

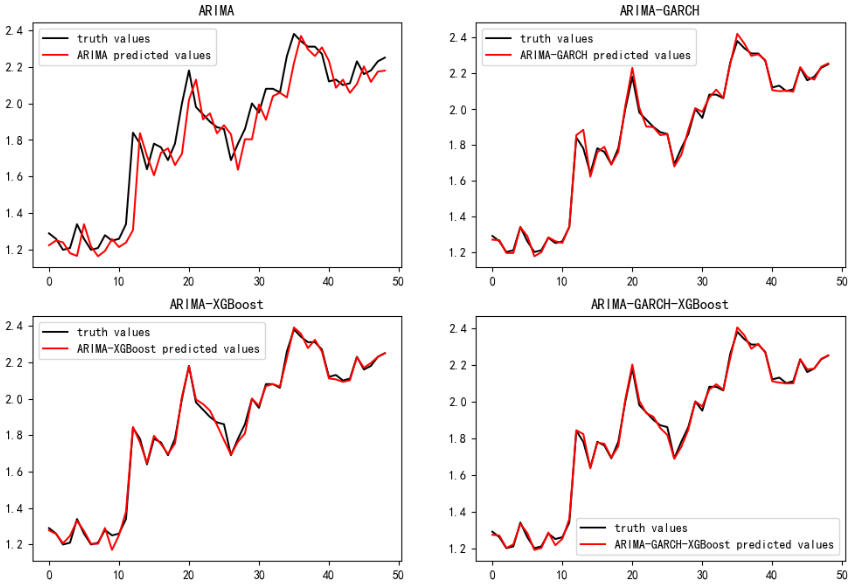

Figure 6: Comparisons of truth values and predicted values

5. Conclusions

This study selects the daily opening price data of Hong Kong stock Sense-W (HK00020) in 2022 as the research subject, and conducts an analysis and prediction of the opening price. Initially, the ARIMA model is utilized for modeling and obtaining predicted values of the opening price. Due to deviations between predicted and actual values from the ARIMA model, GARCH-XGBoost is employed to rectify residual errors, leading to the establishment of an integrated stock opening price prediction model - ARIMA-GARCH-XGBOOST.[16]Empirical findings demonstrate a significant enhancement in predictive performance across three key indicators compared to single-model methods, particularly manifesting a notable increase in determination coefficient from 0.889 to 0.997 with effective residual correction. Moreover, when compared with individual residual correction models such as ARIMA-GARCH and ARIMA-XGBoost, our combined approach not only improves prediction accuracy but also exhibits robustness applicable across diverse stock price prediction tasks. These results not only provide more reliable decision support for investors but also present a valuable integrated model for time series prediction studies. Future endeavors will focus on optimizing residual correction methods to further enhance predictive outcomes while exploring broader applications of our proposed ARIMA-GARCH-XGBOOST combination model.

Acknowledgements

This paper is supported by National Natural Science Foundation of China (No.12461020) and Provincial first-class courses in Gansu Province: Probability Theory and Mathematical Statistics and Discipline Construction Project of Lanzhou City University and Curriculum Ideological and Political Project of Lanzhou City University.

References

[1] Wu Y, Wen X, et al. Short-term stock price prediction based on ARIMA model[J]. Statistics and Decision Making, 2016, 23: 83-86.

[2] Chen Y S, Tang Z J, Luo Y, et al. Research on stock price prediction by Pearson optimization combined with XGBoost algorithm[J]. Information Technology, 2018, 9: 84-89.

[3] Peng Y, Liu Y H, Zhang R F, et al. Modeling and analysis of stock price prediction based on LSTM[J]. Computer Engineering and Applications, 2019, 55: 209-212.

[4] Cao X, Zhao Z. Research on stock index forecasting based on ARIMA-GARCH and SVM mixed model [J]. Proceedings of 2022 International Conference on Software Data Processing and Information Technology, 2022, 1: 41-47.

[5] Feng C, Gao M. An Improved ARIMA Method Based on Functional Principal Component Analysis

and Bidirectional Bootstrap and Its Application to Stock Price Forecasting[J]. Academic Journal of Computing Information Science, 2022, 5: 14-19.

[6] Jin B, Gao S, Tao Z, et al. ARIMA and Facebook Prophet Model in Google Stock Price Prediction[J]. Proceedings of Business and Economic Studies, 2022, 5: 60-66.

[7] Kahkashan K. Price Forecasting of Maruti Suzuki using ARIMA Model[J]. Parikalpana KIIT Journal of Management, 2022, 18: 90-98.

[8] Pratiwi L, Susetyo B, Sadik K, et al. Comparison of Transfer Function Model and ARIMA-GARCH on Daily Stock Data in Agribusiness and Trade Sector[J]. International Journal of Sciences: Basic and Applied Research, 2022, 61: 57-69.

[9] Ying X. Using ARIMA-GARCH Model to Analyze Fluctuation Law of International Oil Price[J]. Mathematical Problems in Engineering, 2022, 22: 11-17.

[10] Ayman A, Saleh B, and Faisal A. Stock price prediction using ARIMA versus XGBoost models: the case of the largest telecommunication company in the Middle East[J]. International Journal of Information Technology, 2023, 15: 1813-1818.

[11] Elangovan R, Gnanasekar F and Parayitam S. Day of the week effect in the Indian stock market[J]. International Journal of International Accounting and Finance, 2023, 11: 181-201.

[12] Goyal R, Adjemian K M, Glauber J, et al. Decomposing USDA Ending Stocks Forecast Errors[J]. Journal of Agricultural and Resource Economics, 2023, 48: 260-276.

[13] Hao Q J. The Application of the ARIMA-GARCH Hybrid Model for Forecasting the Apple Stock Price[J]. Journal of Intelligence and Knowledge Engineering, 2023, 1: 01-08.

[14] Jiancheng S, John G, Mohammad N, et al. Predicting Stock and Bond Market Returns with Emotions: Evidence from Futures Markets[J]. Journal of Behavioral Finance, 2023, 24: 333-344.

[15] Paravee M, Binxiong Z and Woraphon Y. Predicting Chinese stock prices using convertible bond: an evidence-based neural network approach[J]. Asian Journal of Economics and Banking, 2023, 7: 294-309.

[16] Yuan W K, Zhou Z T, et al. Research on the influencing factors of stock price volatility based on PCA-BP neural network[J]. Business and Exhibition Economics, 2023, 10: 99-101.