Goldman Sachs's Price Forecast Based on ARIMA and LSTM

Jiapeng Yang *

School of Social Sciences, University of Manchester, Manchester, The United Kingdom

* Corresponding author: jiapeng.yang@studemt.manchester.ac.uk

Abstract. The prediction of stock prices is a common and crucial problem in trading. Correctly predicting future stock prices enables traders to determine the optimal time to buy and sell stocks, increasing the probability of making profits. This study focuses on predicting the closing price of Goldman Sachs. Initially, an ARIMA (4,1,6) benchmark model was established based on the AIC information criteria for time series prediction. The model was then applied to make forward predictions. Subsequently, a two-layer LSTM model was constructed. The prediction results of both models were visualized, and the regression indicators were calculated for each model. By comparing the prediction results of the two algorithms, it was determined that LSTM model outperforms ARIMA on the dataset used in this paper. Furthermore, this paper highlights some shortcomings of the ARIMA model, including its unsuitability for long-term forecasting and its subjective parameter selection. In contrast, LSTM performs better in predicting turning points in stock prices.

Keywords: Price Forecast, ARIMA, LSTM.

1. Introduction

Forecasting stock prices is a classic problem that has captivated the interest of investors. Investors continually strive to profit from the stock market by accurately forecasting price movements. If stock prices can be effectively predicted, individuals can achieve high returns in the capital markets. Moreover, from an academic standpoint, understanding the predictability of stocks contributes to the study of capital market efficiency and aids researchers in constructing more realistic capital models. Therefore, the issue of stock forecasting holds immense importance. However, the price of stocks is subject to numerous uncertainties, making it challenging for investors to make accurate predictions.

Previous studies on the prediction of the stock prices have predominantly focused on employing single-machine learning models. These include nonlinear time series models [1,2], artificial neural networks [3-5], decision trees [6, 7], genetic algorithms [8, 9], Markov models [10], and support vector machines [11]. However, these previous investigations have certain limitations. To address these gaps, this paper employs both integrated and moving average (ARIMA) [12] and Long Short-Term Memory (LSTM) [13] models as forecasting techniques. The experimental result indicate that LSTM is more accurate in predicting stock price.

The following sections are structured as follows: In Section 2, a thorough examination of ARIMA and LSTM is presented. In Section 3, the dataset utilized in our study is presented. Section 4 outlines the extensive experimental procedures conducted. Finally, some conclusions were drawn in the Section 5 of this paper.

2. Methodology

This section delves into the theoretical foundations, inference procedures, and algorithmic explanations pertaining to ARIMA and LSTM.

2.1. ARIMA

ARIMA is a classic time series analysis algorithm used for modeling and forecasting time series data [14]. It combines the features of autoregressive (AR) and moving average (MA) models and uses differencing to handle non-stationary data. The mathematical representation of the ARIMA is as follows:

$$\Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \tag{1}$$

Where $\Phi(B)$ is defined as:

$$\Phi(B) = 1 - \phi_1 B^1 - \phi_2 B^2 - \cdots - \phi_p B^p \tag{2}$$

∇^d is defined as:

$$\nabla^d = (1 - B)^d \tag{3}$$

$\Theta(B)$ is defined as:

$$\Theta(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - \cdots - \theta_q B^q \tag{4}$$

Where x_t and ε_t are the actual value and random error at time period t, respectively. The parameters p and q represent the orders of the AR and MA components, respectively, while d denotes the number of differencing operations. It's important to note that all of these parameters, p, d and q, are integer values.

$\phi_i(i = 1,2, \ldots, p)$ and $\theta_j(j = 1,2, \ldots, q)$ are parameters of the algorthim. ε_t represents random errors that are assumed to follow an independent and identically distributed (IID) pattern, with a zero mean and a constant variance of σ_ε^2, that is $E(\varepsilon_t) = 0,\ Var(\varepsilon_t) = \sigma_\varepsilon^2$ $,E(\varepsilon_t \varepsilon_s) = 0, s \neq t.$

The AR is composed of $\Phi(B)$, where $\varphi_1, \varphi_2, \ldots, \varphi_\mathrm{p}$ are the autoregressive coefficients that represent the impact of each lagged term on the current observed value. The Integrated part is composed of ∇^d, where d is the order of differencing. Stationary series can be obtained by employing differencing, which transforms non-stationary time series into a more suitable form for analysis and modeling. The MA part is composed of $\Theta(B)$, where $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average coefficients that represent the impact of each lagged term on the current white noise term.

To summarize, the ARIMA model integrates autoregression, differencing, and moving average techniques to effectively model and predict time series data. It is used for analyzing and handling non-stationary time series data.

2.2. LSTM

LSTM is a specific neural network architecture which is to efficiently retain and leverage both long-term and short-term information within sequences of data [15]. It falls under the category of recurrent neural networks (RNNs), but it tackles the challenge of gradient vanishing and exploding that may occur during the training process that arises with traditional RNNs when handling lengthy input sequences. Consequently, RNNs struggle to capture long-term dependencies. LSTM was specifically designed to address this specific limitation and overcome the associated challenges.

In comparison to the hidden state of a standard RNN, LSTM introduces a novel component called the cell state. This cell state serves as the heart of the LSTM architecture. The LSTM network incorporates a mechanism referred to as gates, which allow for selective modification or extraction of information from the cell state. More specifically, LSTM incorporates three distinct gates: forget gates, input gates, and output gates, which together enable the network to control the flow of information.

The forget gate plays a crucial role in filtering out irrelevant information, while the input gate governs mechanism of integrating new information into cell state, and output gate controls the extraction of relevant information from the cell state. By leveraging these gates, LSTM effectively manages long-term dependencies and retains critical information for longer durations.

The forgetting gate is responsible for determining which information should be discarded from the current cell state. By examining the input sequence X_t and the previous hidden state h_{t-1}, the mechanism produces an output vector ranging from 0 to 1, where the range of values between 0 and 1 indicates the extent to which the cell state C_{t-1} retains or discards specific information. The number 0 indicates that no reservation is made, and 1 indicates that all reservations are made. The forget gate is defined as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5}$$

The input gate is responsible for selecting and incorporating new information into the cell state. This particular process is further divided into two distinct stages. First, an operation referred to as an input gate is employed to determine the information to be updated using parameters h_{t-1} and X_t. A tanh layer is applied to h_{t-1} and X_t in order to generate new candidate cell information $\tilde{C}_t$, which can potentially be incorporated into the existing cell information. The input gate is defined as:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{7}$$

Subsequently, the previous cell information C_{t-1} undergoes an updating process to yield the new cell information C_t. This updating procedure involves employing the forget gate to selectively discard a portion of the old cell information, and by utilizing the input gate, a segment of the candidate cell information $\tilde{C}_t$is chosen and incorporated to generate the updated cell information C_t. The forget gate is defined as:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{8}$$

Once the cell state has been updated, it becomes imperative to ascertain the state attributes of the output cell based on the input h_{t-1} and X_t. In this regard, the input is subjected to a sigmoid layer known as the output gate, which yields the decision criteria. Subsequently, the cell state undergoes a transformation through a tanh layer, resulting in a vector with values ranging from -1 to 1. This vector is then multiplied by the decision criteria obtained from the output gate, thereby obtaining the ultimate output of the RNN unit. The output gate is defined as:

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \tag{9}$$

$$h_t = o_t * \tanh(C_t) \tag{10}$$

3. Dataset

Goldman Sachs, founded in 1869, is an esteemed international investment bank, known for being one of the oldest and largest institutions of its kind worldwide [16]. Its long tenure in the market provides a valuable advantage, that is the availability of extensive historical data.

In this study, the Goldman Sachs Investment Company is selected as the research subject, focusing on the time period spanning from 2001 to 2019. The objective is to forecast the final trading price of the Goldman Sachs Investment Company stock, utilizing both the ARIMA and LSTM models. To evaluate the accuracy of these predictions, a designated test set consisting of data from 2006 to 2019 has been employed. By comparing the prediction accuracy and discrepancies between the two models, their respective performance can be assessed. To begin with, a comprehensive examination of the closing price is carried out through descriptive statistical analysis. The findings of this analysis are presented in Table 1. The information presented in Table 1 reveals the price trends observed throughout the study duration, the average closing price of Goldman Sachs Investment Company's stock was 151.8, the observed range of the closing price spans from a minimum value of 52 to a maximum value of 273.38.

Table 1. The statistical summary of the provided data

n	missing	distinct	Info	Mean	Gmd	Median	Min	Max
4778	0	4082	1	151.8	58.95	156.38	52	273.38

4. Experiments

4.1. Evaluation metric

Mean Squared Error (MSE) is a commonly used evaluation metric for regression analysis [17]. It measures the average squared difference between the predicted values of a model and the actual values. A lower MSE indicates higher precision and accuracy in the model's predictions. In the following formulas (11)- formulas (13), y_i is the true value of the i_{th} data point, and the $\hat{y}_i$ is the predicted value of the i_{th} data point, and n is the number of all data points in the dataset. The calculation formula for MSE is as follows:

$$MSE = \frac{1}{n} * \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11}$$

The Root Mean Squared Error (RMSE) is computed by taking the square root of the Mean Squared Error (MSE) [18]. RMSE is commonly used in academic writing as it restores the original scale of the target variable, allowing for a more accurate assessment of the prediction error magnitude. The calculation formula for RMSE is,

$$RMSE = \sqrt{\frac{1}{n} * \sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{12}$$

The Mean Absolute Error (MAE) is determined by calculating the average absolute deviation between the predicted values and the true values [18]. MAE is regarded as a more resistant error metric because it is less affected by outliers or extreme values. The formula for computing MAE is given as,

$$MAE = \frac{1}{n} * \sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{13}$$

4.2. Experimental settings

This paper utilizes the data of Goldman Sachs Investment Company stock from January 2, 2001 to March 14, 2016 as the train dataset for the model. For evaluating the performance of models, the data of Goldman Sachs Investment Company stock from March 15, 2016 to December 30, 2019 is used as the test dataset. The same data processing techniques applied to the training dataset are also employed for the test dataset. As a regression task, the evaluation metrics such as MSE, RMSE, MAE, used in evaluating the model's prediction accuracy for stock prices in the test dataset.

4.3. Results

In this section, the forecasting performance of Goldman Sachs' data will be analyzed in detail using LSTM and ARIMA models, as well as compare the results between the two models.

4.3.1. ARIMA

To assess the model's fitting effectiveness, the time series were partitioned into two segments: the 80% of the data points as training set, and the test set includes the remaining 20% of the data points.

Initially, a stationarity test was performed on the logarithmic closing price. The Augmented Dickey-Fuller (ADF) test produced a statistic of -2.4161, which corresponds to a p-value of 0.4021. Therefore, we fail to reject the null hypothesis, indicating the presence of a unit root in the data. Subsequently, the first-order differencing and performed the unit root test were conducted once again. This time, the ADF test statistic exhibited a value of -14.84, accompanied by a p-value of 0.01. As a result, the null hypothesis was rejected. Consequently, the differenced sequence led to a stationary time series, meeting the requirements for time series modeling.

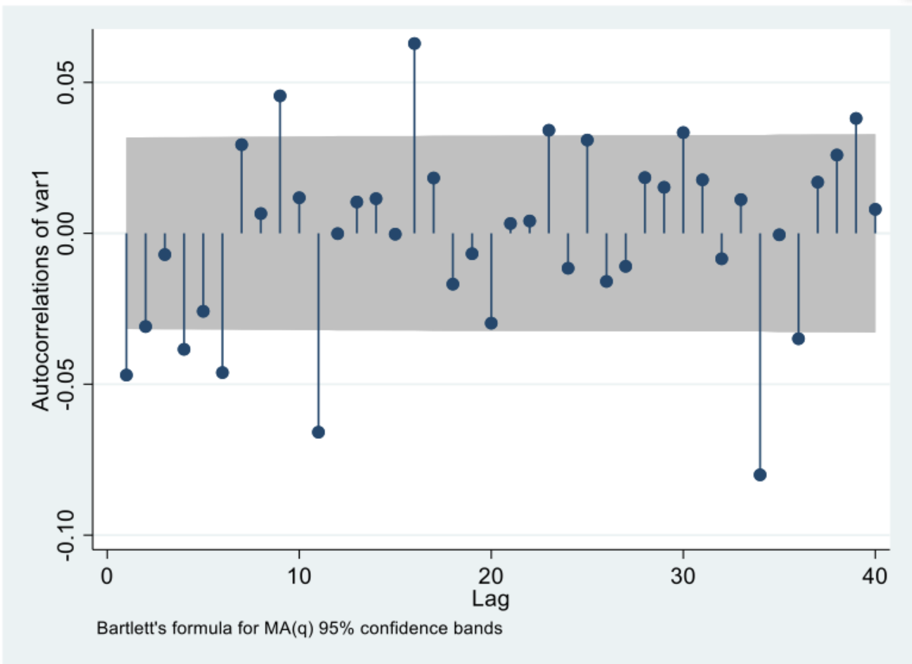

Figure 1. Autocorrelation graph

By applying the aforementioned methods, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) were calculated for determining the appropriate values for parameters p and q [19]. Specifically, for the dataset considered in this study, the ACF was visualized in Fig. 1 and the PACF in Fig. 2. The Figures depict the correlation and partial autocorrelation plots of the time series after performing the first-order differencing. These plots reveal that the differenced time series is not akin to white noise, making it challenging to ascertain the optimal values for parameters p and q. However, employing information criteria, the ARIMA (4, 1, 6) model provided the best fit was concluded. After constructing the model using the train dataset, a white noise test on the residual errors was conducted. The results indicated that the residual errors of the model exhibited characteristics resembling white noise.

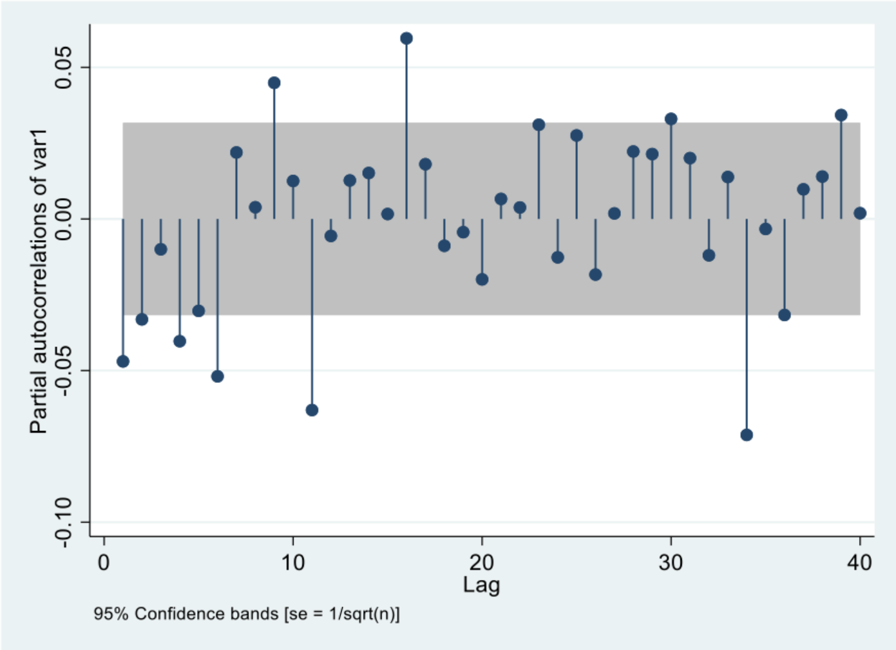

Figure 2. Partial Autocorrelation graph

The white noise test is conducted on a stationary time series, irrespective of the lag period chosen (e.g., 5, 10, or 15). In this analysis, the p-value associated with the test statistic is significantly lower than 0 was observed. Consequently, the initial assumption that the time series conforms to a white noise pattern was rejected. This outcome satisfies the prerequisite for establishing an ARIMA model.

Finally, the model performs forecasts and visualizes the closing prices from 2016 to 2019 (See Fig. 3). In this analysis, A fixed window is employed, and forward rolling prediction is used, where one value is predicted at a time. This approach helps enhance the accuracy of the forecasts. Simultaneously, the RMSE, MAE, and MSE was calculate ed between the predicted and true values. The calculated RMSE, MAE, MSE is 3.1407, 9.8607 and 2.2924, respectively. Based on the results, the predicted values closely align with the true values, indicating a high level of accuracy in the forecasts.

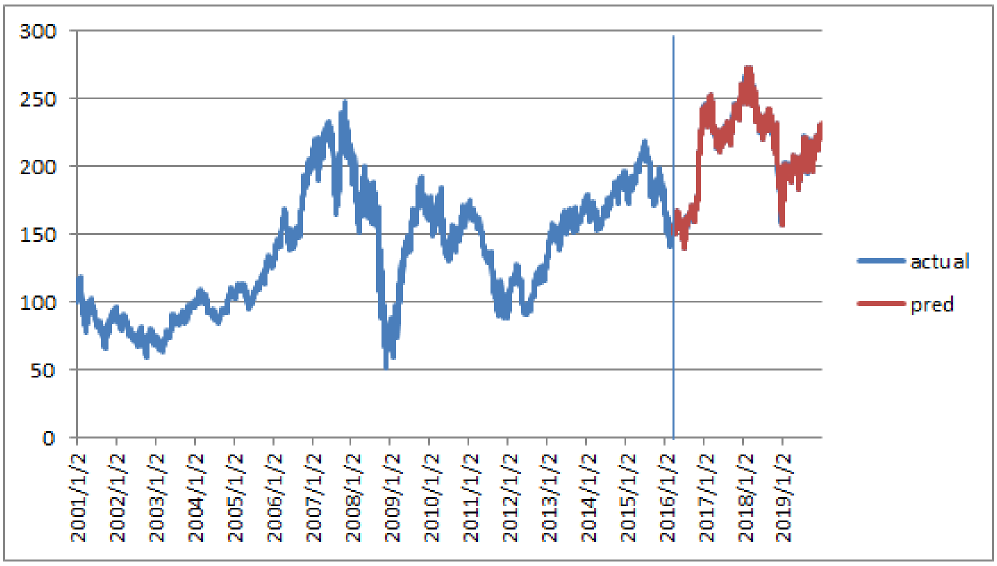

Figure 3. Predictions of the ARIMA model

4.3.2. LSTM

In contrast, a two-layer LSTM model is constructed for prediction and the resulting forecasts are visualized in Fig. 4. Subsequently, the RMSE, MAE, and MSE between the predicted and true values were calculated. The calculated RMSE, MAE, MSE is 3.1407, 7.2355 and 2.0122, respectively. The findings indicate a slight improvement in the prediction performance compared to the ARIMA model.

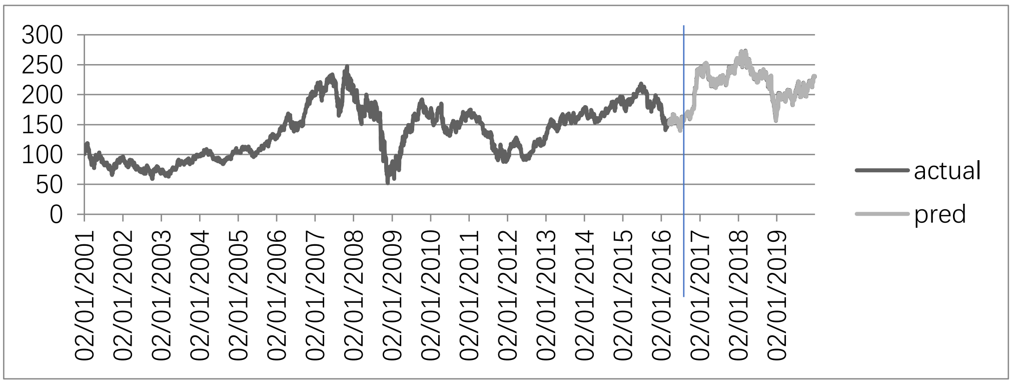

Figure 4. Predictions of the LSTM model

4.3.3. Comparison

Based on the comparison of the forecasting outcomes of the two models, it can be deduced that LSTM outperforms ARIMA in predicting the Goldman Sachs stock price, albeit by a small margin. It is worth noting that ARIMA has certain limitations, as it is not specifically designed for long-term forecasting and its parameters are subjective. The specific comparison details between LSTM and ARIMA are shown in Table 2 and Fig. 5. Moreover, when faced with inflection points or changing trends, ARIMA tends to be less effective in making accurate predictions compared to LSTM.

Table 2. Comparison of LSTM and ARIMA model performance

Model	MSE	RMSE	MAE
LSTM	2.0122	3.0023	7.2355
ARIMA	2.2924	3.1407	9.8607

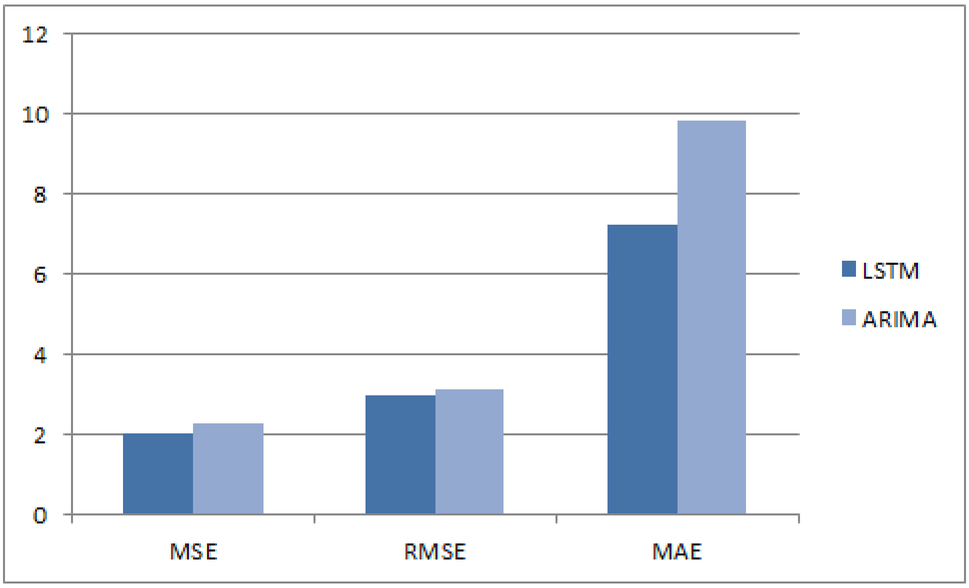

Figure 5. Comparison of Model Results

5. Conclusion

In this paper, ARIMA and LSTM are utilized to forecast the closing price of Goldman Sachs from 2001 to 2019. The prediction performance of these two models is compared and analyzed. MSE, RMSE and MAEare employed as the evaluation metric, and it is found that LSTM outperforms ARIMA for the dataset considered in this study. It is important to note that LSTM has been optimized when compared to other advanced models, thus only LSTM is considered in this paper. However, the exploration of alternative models warrants further investigation.

References

[1] Sezer O B, Gudelek M U, Ozbayoglu A M. Financial time series forecasting with deep learning: A systematic literature review: 2005 – 2019. Applied soft computing, 2020, 90: 106181.

[2] Cheng D, Yang F, Xiang S, et al. financial time series forecasting with multi-modality graph neural network. Pattern Recognition, 2022, 121: 108218.

[3] Mehtab S, Sen J. Analysis and forecasting of financial time series using CNN and LSTM-based deep learning models. Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2021. Springer Singapore, 2022: 405 - 423.

[4] Cao Y, Shao Y, Zhang H. Study on early warning of E-commerce enterprise financial risk based on deep learning algorithm. Electronic Commerce Research, 2022, 22 (1): 21 - 36.

[5] Chen L, Pelger M, Zhu J. Deep learning in asset pricing. Management Science, 2023.

[6] RL M, Mishra A K. Measuring financial performance of Indian manufacturing firms: application of decision tree algorithms. Measuring Business Excellence, 2022, 26 (3): 288 - 307.

[7] Zhu W, Zhang T, Wu Y, et al. Research on optimization of an enterprise financial risk early warning method based on the DS-RF model. International review of financial analysis, 2022, 81: 102140.

[8] Abraham R, Samad M E, Bakhach A M, et al. Forecasting a stock trend using genetic algorithm and random forest. Journal of Risk and Financial Management, 2022, 15 (5): 188.

[9] Goforth E, Yosri A, El-Dakhakhni W, et al. Infrastructure Asset Management System Optimized Configuration: A Genetic Algorithm–Complex Network Theoretic Meta management Approach. Journal of Infrastructure Systems, 2022, 28 (4): 04022029.

[10] Oelschläger L, Adam T. Detecting bearish and bullish markets in financial time series using hierarchical hidden Markov models. Statistical Modelling, 2023, 23 (2): 107 - 126.

[11] Zhao J, Li B. Credit risk assessment of small and medium-sized enterprises in supply chain finance based on SVM and BP neural network. Neural Computing and Applications, 2022: 1 - 12.

[12] Pirani M, Thakkar P, Jivrani P, et al. A comparative analysis of ARIMA, GRU, LSTM and BiLSTM on financial time series forecasting. 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE, 2022: 1 - 6.

[13] Kathirkamanathan N, Thevarasa B, Mahadevan G, et al. Prevention of DDoS attacks targeting financial services using supervised machine learning and stacked LSTM. 2022 IEEE 7th International conference for Convergence in Technology (I2CT). IEEE, 2022: 1 - 5.

[14] Kong L, Li G, Rafique W, et al. Time-aware missing healthcare data prediction based on ARIMA model. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022.

[15] Moghar A, Hamiche M. Stock market prediction using LSTM recurrent neural network. Procedia Computer Science, 2020, 170: 1168 - 1173.

[16] Hudson M. From Marx to Goldman Sachs: the fictions of fictitious capital, and the financialization of industry. Critique, 2010, 38 (3): 419 - 444.

[17] Marmolin H. Subjective MSE measures. IEEE transactions on systems, man, and cybernetics, 1986, 16 (3): 486 - 489.

[18] Chai T, Draxler R R. Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific model development discussions, 2014, 7 (1): 1525 - 1534.

[19] Jain G, Mallick B. A study of time series models ARIMA and ETS. Available at SSRN 2898968, 2017.