S&P 500 ESG Index Prediction with LSTM and ARIMA Model

Jiahui Wu

School of Mathematics and Physics, Xi'an Jiaotong-liverpool University, Suzhou, China

Jiahui.wu18@student.xjtlu.edu.cn

Keywords: ESG, LSTM, ARIMA, Stock index forecast

Abstract: ESG is the abbreviation of environmental, social and corporate governance, which has a positive impact on society, enterprises, regulatory authorities and investors. ESG investment has also attracted more and more attention in recent years. Predicting future ESG stock price trends plays an important role in ESG investment. This paper uses LSTM and ARIMA models to select data of the S&P 500 ESG Index from May 21, 2018, to May 19, 2023 to train the model and make predictions. Finally, MSE and RMSE were used to analyse the accuracy of LSTM and ARIMA model predictions. After analysis, the prediction results of the LSTM model are better than the ARIMA model, and the ARIMA model is simpler to apply.

1. Introduction

ESG stands for Environmental, Social and Corporate Governance, which focuses on the concept of sustainability and environmental protection, and is a non-financial evaluation standard for enterprises. It has a lot of scoring criteria, and the S&P 500 ESG Index is one of them. Businesses with high ratings are more competitive. The ESG evaluation system can provide suggestions for corporate management and investment decisions, help government regulators formulate reasonable policies to regulate corporate behavior, and provide financial institutions with more comprehensive investment decision-making information [2] .

The S&P 500 ESG Index is a stock market index which measures the performance of companies within the S&P 500 universe based on their ESG criteria. ESG investing has gained traction in recent years as investors have become increasingly aware of the importance of sustainability into their investment decisions. The S&P 500 ESG Index can provide investors with a reference to non-financial results.

LSTM and ARIMA models have excellent predictive capabilities for stock prices. Choi [9] employed an ARIMA-LSTM model for predicting stock price correlations, and the results indicated that this hybrid model could enhance the accuracy of predictions. Xiao et al. [10] utilized LSTM and ARIMA models to predict the stock prices of four different companies, and the results showed that LSTM's predictions were better than ARIMA. The LSTM model has a strong predictive ability for the stock index[6].

This study aims to leverage the predictive capabilities of LSTM and ARIMA models to assist investors, financial analysts, and policymakers in making informed decisions regarding ESG investments and portfolio management.

This research utilized LSTM and ARIMA models to model and predict the S&P 500 ESG Index. We collected historical data from May 21, 2018, to May 19, 2023, which was used for model training and forecasting. Following this, in the Methodology section, which explains the principles of the LSTM and ARIMA models. Next, in the Experimental section, we preprocess the data and split it into training and testing sets to build the two models separately. Finally, we compare the prediction results of the two models by calculating MSE and RMSE, and present graphical representations.

2. Literature Review

The price of stocks rises and falls, often irregular, if they can predict future stock prices from past and current stock prices, this is very useful for investors. Predicting the S&P 500 ESG Stock Index

can help ESG investors analyze and judge future stock trends. The LSTM and ARIMA models have good forecasting capabilities and are often used to make predictions about stock prices. Some researchers have successfully predicted the opening price of stocks using the LSTM model and have found it to be more accurate than other models such as RNN and ML [5]. Roondiwala, Patel and Varma [6] use the LSTM method to predict stock prices to help investors and analysts understand where the stock market is headed in the future. In addition, Mondal, Shit and Goswami [7] used the ARIMA model to study 56 stocks in seven industries, and the result was that ARIMA's stock price prediction accuracy was above 85%. The ARIMA model is simple, widely cited, and well suited for short-term forecasting [8].

However, the LSTM and ARIMA models have different advantages and disadvantages. In terms of prediction excellence, the prediction results of the LSTM model are better, while the ARIMA model is simpler than the LSTM [10]. Many researchers have come to the same conclusion, the predictive power of the LSTM model is better than the ARIMA model, but the LSTM model is more affected by data processing [12][11].

We can see that most of the predictions using models for stock prices are based on common stocks, while our study emphasizes ESG stock predictions, hoping to find that LSTM and ARIMA model predictions also apply to ESG stock indices.

Using LSTM and ARIMA models to study the S&P 500 ESG index predicting future stock prices is valuable for studying ESG. And ESG has a positive impact on businesses, investors, government regulators, financial institutions and academic institutions [2]. ESG ratings provide an important aid to companies when raising and investing, identifying and quantifying risks and opportunities beyond finance, while positively impacting a company's sustainability performance [1]. According to research by Alareeni and Hamdan [3], ESG has a Beneficial impact on an enterprise's operations, finances, and markets. ESG information is also very important for investors, and investors are more optimistic about the finances of companies with high ESG[4]. Therefore, the research on predicting ESG stock indices has certain practical significance.

3. Methodology

3.1. LSTM model

- Concept:

LSTM (Long Short-Term Memory) is a special recurrent neural network (RNN) model. In order to retain useful key information, the LSTM model has cell state throughout in addition to the output h of the RNN, which controls how much information passes through one unit and passes to the next unit by adding various gating, such as input gates, forget gates, and output gates.

- Math Model:

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

$$i_t = \sigma(W_i \cdot [b_{t-1}, x_t] + a_i) \tag{2}$$

$$\widetilde{C_t} = \tanb(W_C \cdot [b_{t-1}, x_t] + a_C) \tag{3}$$

$$o_t = \sigma(W_o \cdot [b_{t-1}, x_t] + a_o) \tag{4}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t} \tag{5}$$

$$n_t = o_t \cdot \tanb(C_t) \tag{6}$$

f_t: Recursive connection weights for the forgotten gate
i_t: The input gate's recursive connection weight
o_t: The output gate's recursive connection weight
W: Weight at unit
C: Memory state (cell state)
$\widetilde{C_t}$: Candidate memory state (current moment cell state)
σ (z): Sigmoid activation function

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1+\tanh\left(\frac{z}{2}\right)}{2} \tag{7}$$

tanb(z): tanh activation function

$$tanb(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \tag{8}$$

3.2. ARIMA model

- Concept:

ARIMA model is a differential autoregressive moving model ARIMA (p, d, q) that combines autoregressive model (AR), moving average model (MA) and differential method, which is commonly used in time series predictive analysis.

- Math Model:

$$\left(1-\sum_{i=1}^{p}\beta_i L^i\right)(1-L)^d X_t = \left(1+\sum_{i=1}^{q}\beta_i L^i\right)\varepsilon_t \tag{9}$$

p is the number of autoregressive terms; q is the number of terms of the sliding average; d is the number of differences (order), L is Lag operator.

4. Experimental

4.1. LSTM

4.1.1. Visualize

To have an initial observation of the data, it is necessary to calculate the moving average using rolling windows of sizes 10, 20, 30, and 40, respectively, and then plot multiple line charts. Figure 1 depicts the time series smoothed by the moving averages, enabling a clearer visualization of the stock price trends.

The final sentence of a caption must end with a period.

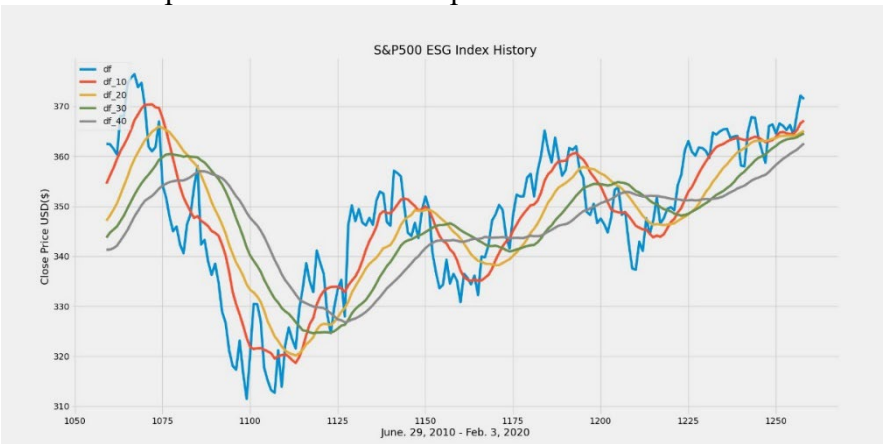

Figure 1 Line chart of the moving average of the ESG index's closing price.

4.1.2. Normalization

Scale the data into a uniform range of 0 to 1. The formula of Min-Max Normalization is:

$$Normalized\ value = \frac{Original\ value\ -\ Minimum\ value}{Maximum\ value\ -\ Minimum\ value} \tag{10}$$

4.1.3. Training Setting

Use 80% of the data as training data and 20% as test data. With a sample size of 948, set the time step to 60 and construct an LSTM model. The loss function is Mean Squared Error (MSE):

$$MSE = \frac{1}{n} * \Sigma(y - \hat{y})^2 \quad (11)$$

n is the total number of samples or data points.

y is the actual values of ESG close price.

ŷ is the predicted values of ESG close price.

In this study, the model is trained by optimizing the deviation between actual values and predicted values through adjusting the MSE values of the data. Additionally, ADAM is used as an optimizer to more effectively update model parameters during the training process, making it particularly suitable for problems with large datasets and parameters.

4.2. ARIMA

4.2.1. Stationarity

Before building an ARIMA model, it is important to ensure that the data is stationary. Stationary means that the data is stable and inertial on the time series. The ADF test can be used to determine whether the data is stationary. If it is not, we need to perform a first-order differential and then perform an ADF test on the first-order differential data. If it is still not stationary, you need to continue with second-order differential until the data has stationarity.

The ADF test results of this study showed that the test statistic was -1.1510784168061743, which was greater than the critical value when the significance level was 5%, and the p-value was 0.6942788803198837, which was greater than 0.05. Therefore, not negating the null hypothesis indicates that the data are not stationary. Next, first-order differential is performed on the time serial data, followed by an ADF test on the differential data. The results show that the test statistic is -10.96990330032795, which is less than the critical value when the significance level is estimated by the 5% parameter. (See Figure 2 and 3)

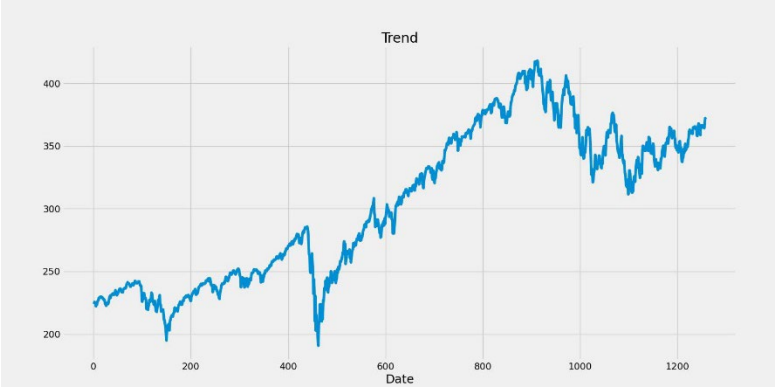

Figure 2 Close price line chart.

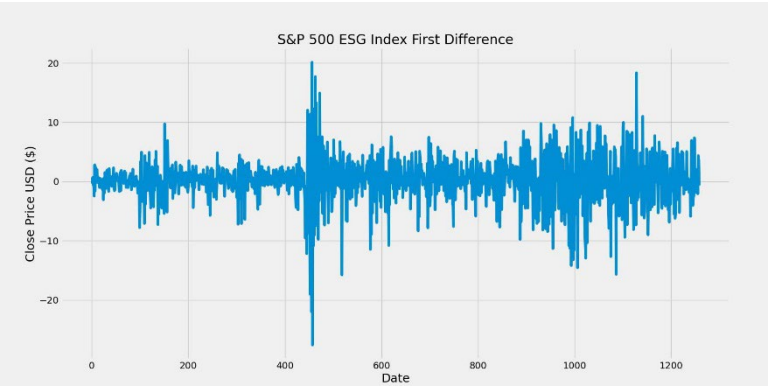

Figure 3 S&P500 ESG index first difference

4.2.2. Parameter estimation

To build an ARIMA model, three parameters need to be determined: p, d, and q. p is the

autoregressive (AR) order. d is the differencing (D) order. q is the moving average (MA) order. Time series data is first-order difference, so d=1. Then, the number of p should be truncated from the PACF plot, while the value of q depends on the truncation of the ACF plot. (See Figure 4)

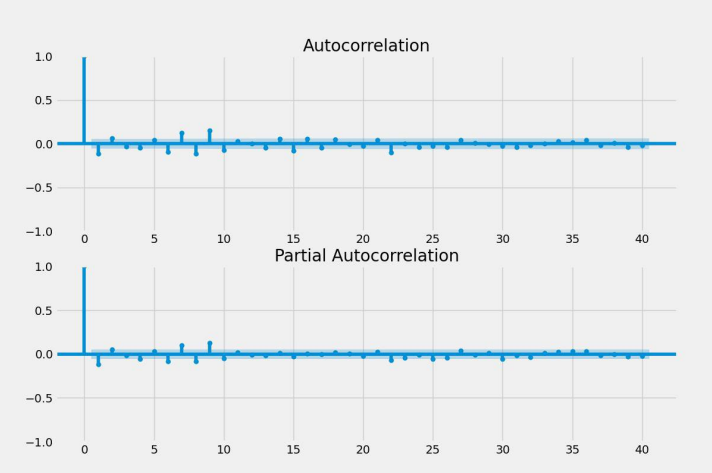

Figure 4 ACF, PACF plots.

From the ACF plot, it can be observed that q=1, indicating a significant correlation at lag 1. From the PACF plot, it can be observed that p=1, indicating a significant partial autocorrelation at lag 1. Therefore, the final model established in this paper is ARIMA (1,1,1).

4.2.3. Residual analysis

Plotting the residual graph reveals that it has a similar trend to the first-order difference plot. Furthermore, examining the kernel density estimation plot of the residuals shows a distribution that closely resembles a normal distribution. The graph is symmetric around a mean of zero. Then describe the residuals, and statistics can see the mean, standard deviation, maximum, minimum, and quartile. The residual test results show that the model fits well. (See Figure 5 and 6)

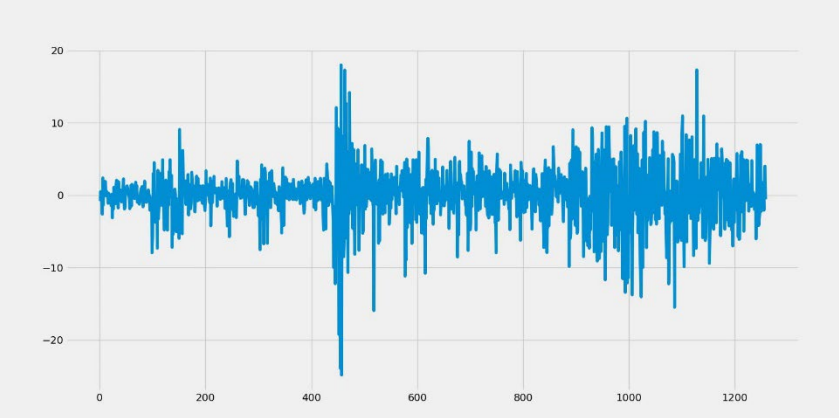

Figure 5 Residual plot.

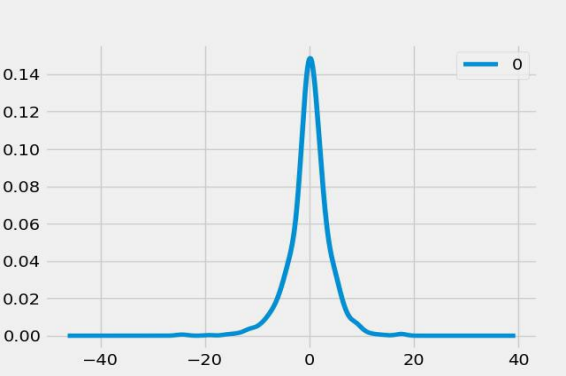

Figure 6 Residual kernel density estimation plot.

Finally, the ARIMA model can be used to make predictions on the time series data.

5. Results and Conclusions

5.1. Prediction of LSTM model

To analyze the prediction accuracy of the LSTM model, we calculate the values of MSE and RMSE, MSE equal to 10.731991274590348 and RMSE equal to 3.2759718061348373. Then draw a line chart. (See Figure 7)

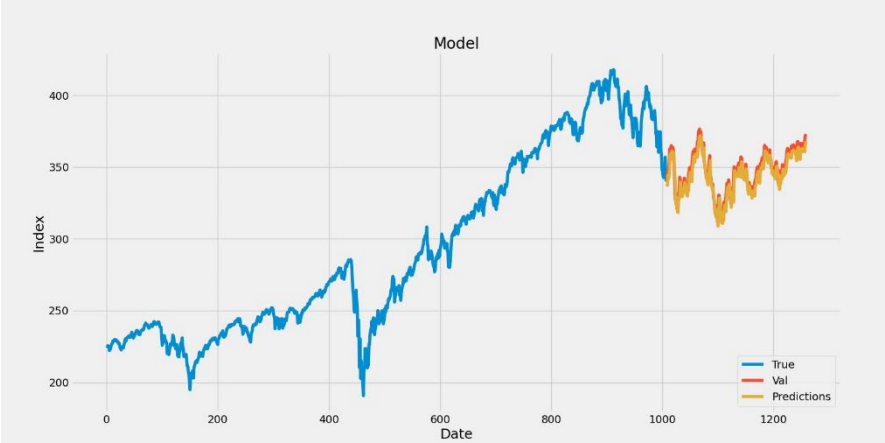

Figure 7 S&P500 ESG index prediction plot of the LSTM model.

5.2. Prediction of ARIMA model

From the experimental section, it can be observed that this study uses an ARIMA (1,1,1) model. By training the model with the first 80% of the time series, predictions are made for the remaining 20% of the data. We compare the differences between the model's predicted results and the actual prices by calculating MSE as Equation (11) and RMSE is obtained by using the MSE rooting number. The MSE is calculated to be 34.246, and the RMSE is 5.852. Then, plotting a line graph to visually display the predicted trend of the model. (See Figure 8)

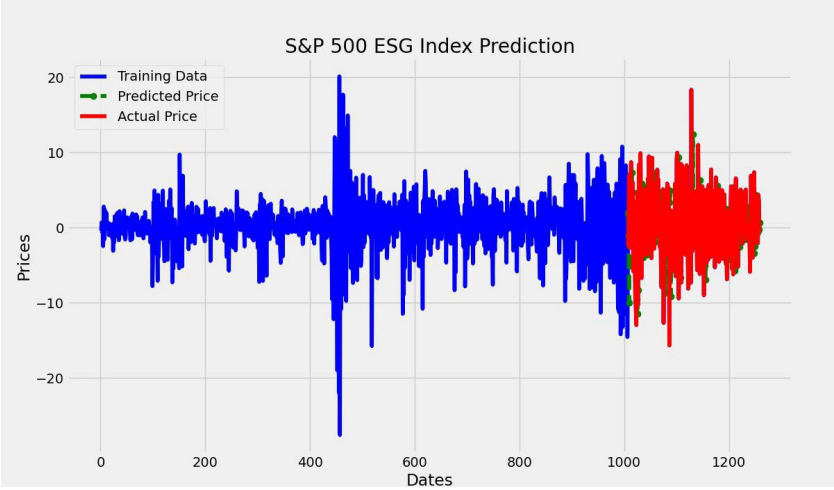

Figure 8 S&P500 ESG index prediction plot of the ARIMA model.

5.3. Comparative analysis of ARIMA and LSTM

The MSE and RMSE values of the LSTM model are smaller than the values of the ARIMA model, indicating that the predictions of LSTM are closer to the real stock price. This suggests that the LSTM model may be more effective in capturing the underlying patterns and trends in the data. (See Table 1)

Table 1 MSE and RMSE comparison of LSTM and ARIMA models.

	LSTM	ARIMA
MSE	10.73199	34.246
RMSE	3.27597	5.852

5.4. Conclusions

Both the LSTM and ARIMA models demonstrate good predictive performance for the S&P ESG stock index. However, the LSTM model exhibits superior forecasting capabilities. It is able to capture more complex patterns and dependencies in the data, allowing it to provide more accurate and reliable predictions compared to the ARIMA model. The reason for this may be that the ARIMA model is only suitable for linear time series and is more sensitive to nonlinearity and outliers. However, the LSTM model requires a large amount of data and computational resources, and parameter tuning can be more complex. On the other hand, the ARIMA model has a relatively simple principle, fast training speed, and strong interpretability.

From another point of view, stock price fluctuations are not only related to time changes, but also to ESG ratings, political factors, and other stocks. The LSTM model and ARIMA model need to be better optimized and applied.

References

[1] D' Amato, V., D' Ecclesia, R., and Levantesi, S. (2022) 'ESG score prediction through random forest algorithm', Computational Management Science, 19(2), pp. 347–373. doi:10.1007/s10287-021-00419-3.

[2] Li, T., Wang, K., Sueyoshi, T. and Wang D. D. (2021) 'ESG: Research Progress and Future Prospects', Sustainability, 13(11663), p. 11663. doi:10.3390/su132111663.

[3] Alareeni, B.A. and Hamdan, A. (2020) 'ESG impact on performance of US S&P 500-listed firms', Corporate Governance: The International Journal of Business in Society, 20(7), pp. 1409–1428. doi:10.1108/CG-06-2020-0258.

[4] Amel-Zadeh, A. and Serafeim, G. (2018) 'Why and How Investors Use ESG Information: Evidence from a Global Survey', Financial Analysts Journal, 74(3), pp. 87–103. doi:10.2469/faj.v74.n3.2.

[5] Zhuge, Q., Xu, L. and Zhang, G. (2017) 'LSTM Neural Network with Emotional Analysis for Prediction of Stock Price', Engineering Letters, 25(2), pp. 64–72.

[6] Roondiwala, M., Patel, H. and Varma, S. (2017) 'Predicting Stock Prices Using LSTM', International Journal of Science, 6(4), pp.1754-1756. doi:10.21275/art20172755.

[7] Mondal, P., Shit, L and Goswami, S. (2014) 'Study of effectiveness of time series modelling (ARIMA) in forecasting stock price', International Journal of Computer Science, Engineering and Applications, 4(2), pp. 13-29. doi: 10.5121/ijcsea. 2014.4202

[8] Khandelwal, S. and Mohanty, D. (2021) 'Stock Price Prediction Using Arima Model', International Journal of Marketing & Human Resource Research, 2(2), pp.98-107.

[9] Choi, H.K. (2018) 'Stock Price Correlation Coefficient Prediction with ARIMA-LSTM Hybrid Model'. Available at: https://search-ebscohost-com-s.elink.xjtlu.edu.cn:443/login.aspx?direct=true&db=edsarx&AN=edsarx.1808.01560&site=eds-live&scope=site (Accessed: 2 July 2023).

[10] Xiao, R., Feng, Y., Yan, L. and Ma, Y. (2022) 'Predict stock prices with ARIMA and LSTM'. Available at: https://search-ebscohost-com-s.elink.xjtlu.edu.cn:443/login.aspx?direct=true&db=edsarx&AN=edsarx.2209.02407&site=eds-live&scope=site (Accessed: 2 July 2023).

[11] Siami-Namini, S., Tavakoli, N. and Namin, A. S. (2018) 'A Comparison of ARIMA and LSTM in Forecasting Time Series', 2018 17th IEEE International Conference on Machine Learning and

Applications (ICMLA), Machine Learning and Applications (ICMLA), 2018 17th IEEE International Conference on, ICMLA, pp. 1394–1401. doi:10.1109/ICMLA.2018.00227.

[12] Ma, Q. (2020) 'Comparison of ARIMA, ANN and LSTM for Stock Price Prediction', E3S Web of Conferences, 218, p. 01026. doi:10.1051/e3sconf/202021801