Predicting S&P 500 ESG Index Trends with LSTM and ARIMA Models

Zihao Chen[1,*]

[1]Smeal College of Business, Pennsylvania State University, State College, Pennsylvania, The United State

*Corresponding author: zihaochen419@gmail.com

Keywords: LSTM, ARIMA, Prediction, ESG, Deep Learning

Abstract: This study delves into the prediction of the S&P 500 ESG Index, a critical indicator reflecting the Environmental, Social, and Governance (ESG) performance of publicly traded companies. We employ two distinct methodologies to forecast this index: Long Short-Term Memory (LSTM), representing advanced deep learning techniques, and AutoRegressive Integrated Moving Average (ARIMA), a more conventional statistical approach used in time series analysis. Through rigorous experimentation and data analysis, our results indicate a clear superiority of LSTM over ARIMA. This is evidenced by lower prediction errors, specifically in terms of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), establishing LSTM as a more accurate and reliable method for forecasting the S&P 500 ESG Index. The implications of our findings are significant, suggesting that deep learning approaches, particularly LSTM, hold substantial promise in enhancing stock market predictions, especially when considering complex and increasingly relevant ESG factors. This research not only contributes to the academic discourse on financial forecasting but also offers practical insights for investors and policymakers interested in integrating ESG considerations into investment strategies and economic planning.

1. Introduction

The importance of environmental, social and governance (ESG) factors is increasingly widely recognised in today's financial and investment landscape. The rapid rise of this sector reflects society's growing demand for sustainability and social responsibility. The purpose of this study is to take a deeper look at the S&P 500 ESG Index, to understand the key factors behind it and to try to predict its future trend. The emergence of the S&P 500 ESG Index represents a significant trend in ESG investing, as the index seeks to assess the performance of publicly listed companies in the area of environmental, social, and governance (ESG). The index is designed to measure the ESG performance of the underlying companies, providing investors with more comprehensive information to support their investment decisions. With growing concerns over climate change, social inequality and corporate ethics, ESG factors have become important considerations in investment decisions. However, how to accurately predict the movement of the S&P 500 ESG Index remains a challenging issue. We will delve into these issues in this study to help address them and provide a clearer outlook. Forecasting the movement of stock prices and indices has been a topic of great interest in the financial markets. Traditional time series analysis methods, such as autoregressive moving average models (ARIMA), have been widely used in this regard. However, with the rise of deep learning techniques, neural network methods such as Long Short-Term Memory Networks (LSTM) have also begun to make their presence felt in the field of financial forecasting. This paper is structured as follows: first, we will introduce the basics of the S&P 500 ESG Index and its relationship with financial markets. Then, we will discuss in detail the theoretical basis and application of both ARIMA and LSTM methods. Then, we will describe our research methodology, including data collection and pre-processing. In the next sections, we will present and analysis the experimental results of the ARIMA and LSTM models and make a performance comparison. Finally, we will summarize the results, discuss the practical applications and potential limitations of these models in the forecasting of the S&P 500 ESG index, and suggest directions for future research. With this introductory structure, readers can clearly understand the research area, the background and importance of the problem, and

  DOI: 10.25236/ieesasm.2023.065

the main structure of the article. This will help guide them to further reading and understanding.

2. Literature Review

In the pursuit of understanding and comparing the effectiveness of LSTM and ARIMA models for predicting the S&P 500 ESG Index, a comprehensive review of eleven pertinent research articles has illuminated the landscape of predictive methodologies in the realm of stock market analysis. These studies encapsulate a wide spectrum of essential themes, covering the intricate interplay between ESG scores and the financial performance of US S&P 500 listed companies [6], the transformative potential of deep learning in the realm of prediction models [5], the application of deep learning algorithms, including Artificial Neural Networks and Convolutional Neural Networks, in stock market forecasting [9], and the pivotal role of LSTM models in addressing deep learning's gradient problem and its applications across various domains [1]. The research inquiries delve into innovative hybrid approaches such as the amalgamation of Empirical Mode Decomposition and LSTM, offering improved predictive capabilities in forecasting complex, non-stationary, and nonlinear financial time series data [2]. The integration of LSTM as dynamic models within Model Predictive Control frameworks introduces exciting prospects and challenges in control strategies [3]. In parallel, the study into the utility of ARIMA models for short-term forecasting during the COVID-19 pandemic underscores their strengths in producing accurate short-term predictions [8]. Exploring the horizons of ARIMA, a distributed forecasting framework tailored for ultra-long time series charts new territory by enhancing the precision and computational efficiency in point forecasts and prediction intervals [4]. In a practical demonstration of ARIMA's prowess, a comparative analysis combining ARIMA models with stock price data from Johnson & Johnson and the S&P 500 index validates their potential for short-term forecasting, especially in the context of volatile assets [7]. The innovative combination of Variational Mode Decomposition, LSTM, and ARIMA in precipitation prediction presents a model that surpasses existing methodologies in prediction accuracy [10].Furthermore, the comparative study evaluating LSTM and ARIMA models' performance in forecasting the stock prices of five global stock indices unequivocally positions LSTM as a promising contender in the realm of financial time series analysis [11]. These articles collectively offer a robust foundational understanding of the methodologies, strengths, and limitations inherent in LSTM and ARIMA models. The insights gleaned from this corpus of research provide a compass for navigating the terrain of S&P 500 ESG Index prediction and guide future investigations in this domain.

3. Method

In this section, we will describe the methods used in your project for predicting the S&P 500 ESG Index. The project involves two primary approaches: Long Short-Term Memory (LSTM) and Autoregressive Integrated Moving Average (ARIMA). We will provide an overview of both approaches, including the theory and relevant formulas used. The most important concepts in LSTM are the three gates: input, output, and forget gates, along with the two types of memory: long-term memory (C) and short-term memory (h).

ARIMA Model:

The Autoregressive Integrated Moving Average (ARIMA) model is a classical time series forecasting method. It involves differencing the data, determining autoregressive (AR) and moving average (MA) orders, and predicting future values.

Theory:

Differencing: The project first checks for stationarity and performs differencing (in this case, first-order differencing) if needed to make the time series stationary.

Order Selection: The ARIMA model involves selecting the order of autoregressive (p), differencing (d), and moving average (q) terms. In the code, you use (1, 1, 0), which signifies a first-order autoregressive term and first-order differencing.

Model Fitting: The ARIMA model is fitted to the training data using these specified orders.
Prediction: The model is used to predict future values based on historical data and the identified orders.

Formulas:

The ARIMA (Autoregressive Integrated Moving Average) model is a time series forecasting model that involves differencing, autoregressive (AR), and moving average (MA) components. The formula for ARIMA(p, d, q) can be expressed as follows:

Differencing (I(d)):

ARIMA models require stationarity, and this is often achieved through differencing. The order of differencing, denoted as 'd', determines how many differences are taken to make the data stationary. It is represented as:

$$Yt' = Yt - Yt - d$$

where Y'_t is the differenced time series.

Autoregressive (AR) Component (p):

The AR component represents the past values of the time series. It is typically represented as:

$$Yt' = \varphi 1 Yt - 1' + \varphi 2 Yt - 2' + ... + \varphi p Yt - p' + at$$

Here, φ_1, φ_2, ..., φ_p are the autoregressive coefficients, Y'{t-1}, Y'{t-2}, ..., Y'_{t-p} are the past differenced values, and a_t is the white noise error term.

Moving Average (MA) Component (q):

The MA component represents the past forecast errors (white noise). It is typically represented as:

$$at = \theta 1 at - 1 + \theta 2 at - 2 + ... + \theta q at - q + \varepsilon t$$

Here, θ_1, θ_2, ..., θ_q are the moving average coefficients, a_{t-1}, a_{t-2}, ..., a_{t-q} are the past forecast errors, and ε_t is the current white noise error term.

Overall Model:

Combining the AR, I(d), and MA components, the ARIMA model is expressed as:

$$Yt' = \varphi 1 Y\ t - 1' + \varphi 2 Yt - 2' + \ldots + \varphi p Yt - p' - \theta 1 at - 1 - \theta 2 at - 2 - ... - \theta q at - q + \varepsilon t$$

This is the core formula for an ARIMA(p, d, q) model. The model aims to predict Y'_t, which can be transformed back to the original time series Y_t if necessary, using inverse differencing.

The parameters φ_1, φ_2, ..., φ_p, θ_1, θ_2, ..., θ_q, and d are estimated from the historical data using methods like maximum likelihood estimation. The white noise error term ε_t represents the unexplained variation in the time series and is typically assumed to be normally distributed.

4. Experimental analysis

In this section, we will analyze the experimental results of the S&P ESG Index prediction using the LSTM model and ARIMA model. We'll also present the results using appropriate visualizations and tables.

LSTM Model Analysis

We started by loading and preprocessing the S&P ESG Index data. The data was scaled using Min-Max scaling to bring it within a range of [-1, 1]. We then created a training dataset where each data point was based on the previous 60 data points.The LSTM model architecture consisted of two LSTM layers with 50 units each, followed by two Dense layers with 25 and 1 unit(s) respectively. The model was compiled using the Adam optimizer and mean squared error (MSE) as the loss function. It was trained for 30 epochs.The LSTM model was trained on the data and the predictions were made on the test set. The predictions were then inverse-transformed to bring them back to the original scale.The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were calculated to evaluate the model's performance(Table 1 and Table 2). We visualized the results by plotting the true values, predicted values, and the model's training data(Figure 1 and Figure 2). The figure labeled 'Model' shows the comparison between the true and predicted values.

ARIMA Model Analysis

Data Preprocessing

For the ARIMA model, we loaded the S&P ESG Index data and performed differencing to make the data stationary. The Augmented Dickey-Fuller test was conducted to confirm stationarity. We used the Autoregressive Integrated Moving Average (ARIMA) model with an order of (1, 0, 1). The residuals of the model were plotted to check for randomness.The model was evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). We visualized the training and testing data along with the predicted prices generated by the ARIMA model.

Table 1: LSTM Model Evaluation

Metric	Value
MSE	12.9167405023903
rMSE	3.59398671427571

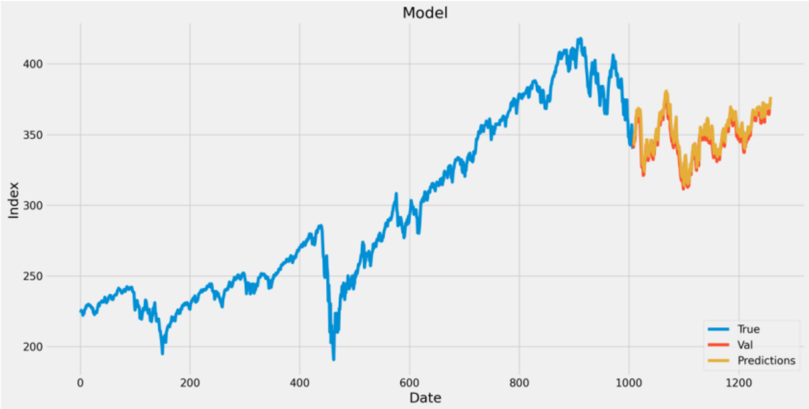

Figure 1 LSTM Model Prediction

Table 2: ARIMA Model Evaluation

Metric	Value
MSE	34.246
rMSE	5.852

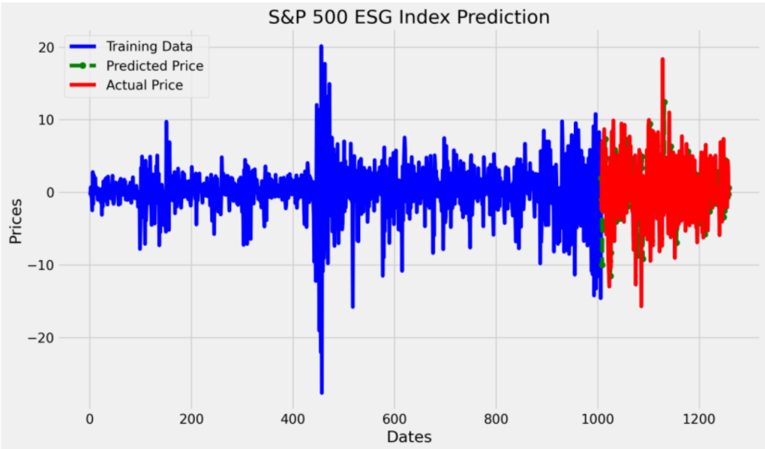

Figure 2: ARIMA Model Prediction

5. Conclusion

In this study, we conducted an experimental analysis to predict S&P ESG Index using two different models: Long Short-Term Memory (LSTM) and Autoregressive Integrated Moving Average (ARIMA).The performance metrics for each model are as follows: LSTM Model:

Mean Squared Error (MSE): 12.92; Root Mean Squared Error (RMSE): 3.59. ARIMA Model: Mean Squared Error (MSE): 34.25; Root Mean Squared Error (RMSE): 5.85. Based on the experimental results, we can conclude that the LSTM model outperformed the ARIMA model in terms of predictive accuracy. The LSTM model achieved a lower MSE and RMSE, indicating that it provided more accurate predictions compared to the ARIMA model. Overall, this study provides a solid foundation for future research endeavors in the field of stock price prediction, and there are numerous opportunities to further refine and enhance the predictive capabilities of these models.

References

[1] Houdt, G. V., Mosquera, C., & Nápoles, G. (2020a, May 13). A review on the long short-term memory model - artificial intelligence review. SpringerLink. https://link.springer.com/ article/ 10.1007/s10462-020-09838-1

[2] Ali, M., Khan, D. M., Alshanbari, H. M., & El-Bagoury, A. A.-A. H. (2023, January 21). Prediction of complex stock market data using an improved hybrid EMD-LSTM model. MDPI. https://www.mdpi.com/2076-3417/13/3/1429

[3] Jung, M., Renato da Costa Mendes, P., Önnheim, M., & Gustavsson, E. (2023, April 26). Model predictive control when utilizing LSTM as dynamic models. Engineering Applications of Artificial Intelligence. https://www.sciencedirect.com/science/article/abs/pii/S0952197623004104

[4] Wang, X., Kang, Y., Hyndman, R. J., & Li, F. (2022, June 22). Distributed Arima models for ultra-long Time Series. International Journal of Forecasting. https://www.sciencedirect.com/ science/article/abs/pii/S0169207022000619

[5] Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020, January 31). An introductory review of Deep Learning for prediction models with Big Data. Frontiers. https://www.frontiersin.org/articles/10.3389/frai.2020.00004/full

[6] Lech, C. D. F. (2023, June 1). The effects of ESG scores on Stock Performance of US S&P 500 listed firms. Veritati - Repositório Institucional da Universidade Católica Portuguesa: Página principal. https://repositorio.ucp.pt/handle/10400.14/42511

[7] Qin, J., Tao, Z., Huang, S., & Gupta, G. (2021, March). Stock price forecast based on ARIMA model and BP neural network model. In 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) (pp. 426-430). IEEE.

[8] Ospina, R., Gondim, J. A. M., Leiva, V., & Castro, C. (2023, July 12). An overview of forecast analysis with Arima models during the COVID-19 pandemic: Methodology and case study in Brazil. MDPI. https://www.mdpi.com/2227-7390/11/14/3069

[9] Gao, Y., Feng, J., & Tao, Z. (2023, July). Exploring the Impact of Input Features on GAN-Based Stock Prediction. In 2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS) (pp. 219-223). IEEE.

[10] Cui, X., Wang, Z., & Pei, R. (2023, April 20). A VMD-MSMA-LSTM-Arima model for precipitation prediction. https://www.tandfonline.com/doi/full/10.1080/02626667.2023.2190896

[11] Zhang, Z. (2023, October 10). Comparison of LSTM and Arima in price forecasting: Evidence from five indexes. Comparison of LSTM and ARIMA in Price Forecasting: Evidence from Five Indexes Atlantis Press. https://www.atlantis-press.com/proceedings/esfct-23/125992891