Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/57/20230772

Comparison of ARIMA and LSTM in Different Industries

Haoxuan Li [1,a,*]

[1]*College of Management, Shenzhen University, Shenzhen, China*
a. 2020041058@email.szu.edu.zn
**corresponding author*

Abstract: In the realm of finance, stock price prediction holds significant importance, and the breakthroughs in deep learning have provided new solutions to this problem. This study compares deep learning methods with traditional time series models by employing a time series model, namely Arima, and the neural network LSTM model to predict the closing stock data of Pfizer Inc. (PFE) in the healthcare sector, Alibaba Group Holding Limited in the e-commerce sector, EA Sports in the gaming industry, and NVIDIA Corporation in the AI industry over the past year. The aim is to identify the most suitable prediction method for each company in stock price forecasting. Through the comparison, it is observed that, regardless of the company, ARIMA outperforms LSTM in stock price prediction. Therefore, it can be concluded that, among these two models, ARIMA is a better fit for stock price prediction in these four companies. Furthermore, it is inferred that this method may also be applicable to other companies within the respective industries. However, further research and data collection are necessary to validate and support this inference. Overall, this study demonstrates the superiority of the ARIMA model over LSTM in stock price prediction for the selected companies in the healthcare, e-commerce, gaming, and AI industries.

Keywords: Arima, LSTM, MSE, stock price predictions

1. Introduction

With the progress of society and the development of human civilization, there is an increasing focus on the returns provided by financial assets. Whether it is individual investors or institutional investors, they aspire to accurately predict the fluctuation of asset prices. This is particularly evident in stock prediction, as precise stock forecasting provides investors with a basis for investment decision-making, enabling them to formulate wiser buying and selling strategies. Therefore, for investors, mastering effective methods and techniques for stock prediction is the key to achieving investment success.

Before the widespread adoption of computers, trading in stocks and commodities was largely driven by individuals' intuition, lacking a scientific and systematic approach to forecasting. As the level of investment and trading grew, attention turned towards the search for tools and methods which could enhance gains while mitigating risks [1]. According to Dutta's study in 2012, which focused on the Indian stock market and employed logistic regression, the model demonstrated favorable performance in assisting investors in selecting high-performing stocks [2]. Over time, an increasing number of statistical models have been introduced. For instance, Zhong and Enke's research in 2017 utilized time series as input variables and encompassed a range of predictive models, including the

DOI: 10.54254/2754-1169/57/20230772

Arima, the Auto-Regressive Moving Average, the GARCH volatility model and STAR model [3]. The ARIMA model combines AR and MA models [4]. It is highly esteemed as an exceptionally efficient forecasting technique in the realm of social science and finds extensive applications [5]. However, the ARIMA model still exhibits some significant limitations. For instance, it does not account for volatility clustering and has a certain dependency on intervention analysis. These factors necessitate cautious consideration when utilizing the model for stock forecasting purposes [6]. Over the past two decades, machine learning models have garnered significant attention [7]. For instance, Patel conducted a comparative analysis of the SVM, random forest, ANN and naive-Bayes models for stock market prediction in the context of Indian markets [8]. All of this has been made possible due to Lapedes and Farber demonstrated the successful application of ANN in nonlinear time series models through their research [9]. Subsequently, other models emerged, including decision trees, and nearest neighbor regression [10-11]. Despite extensive research in these fields, there is limited comparative study of various methods for stock prediction, and a unified standard has not yet been established.

This study will focus on the comparative suitability of two different models, Arima and LSTM, for stock prediction in various industries, and the core goal is to identify the best-fitting predictive model in each domain. Based on Kevin Morgan's research on sunrise industries, namely healthcare, e-commerce, gaming, and AI, these four industries are considered as varying degrees of emerging industries. Thus, this paper will randomly select one major company from each domain, namely Pfizer Inc (PFE) for healthcare, Alibaba Group Holding Limited for e-commerce, EA Sports for gaming, and NVIDIA Corporation for AI. The study will observe their stock data on the NASDAQ from August 3, 2022, to August 2, 2023. The dataset prior to May 2, 2023, will be used as the training dataset, while the subsequent data will serve as the testing dataset. By separately building Arima and LSTM models to fit and predict the training data and comparing them with the testing dataset, the corresponding Mean Squared Error (MSE) will be obtained. The final conclusion suggests that the ARIMA model demonstrates superior performance compared to LSTM in forecasting the closing stock prices of the selected four companies over the previous year. Therefore, it can be considered that, under the chosen parameters for this study, utilizing ARIMA is more suitable for forecasting the stock prices of these four companies.

2. Dataset and Methodology

2.1. Dataset

The data for this study was sourced from Yahoo Finance. This study selected the data for Pfizer Inc. (PFE), Alibaba Group Holding Limited, EA Sports, and NVIDIA Corporation for the past year and further divided the data into a 9-month training set and a 3-month testing set. Non-trading days were excluded from the dataset, resulting in 187 valid data samples for the testing set and 64 data samples for the training set, with no missing values. Some basic information is shown in Figure 1.

DOI: 10.54254/2754-1169/57/20230772

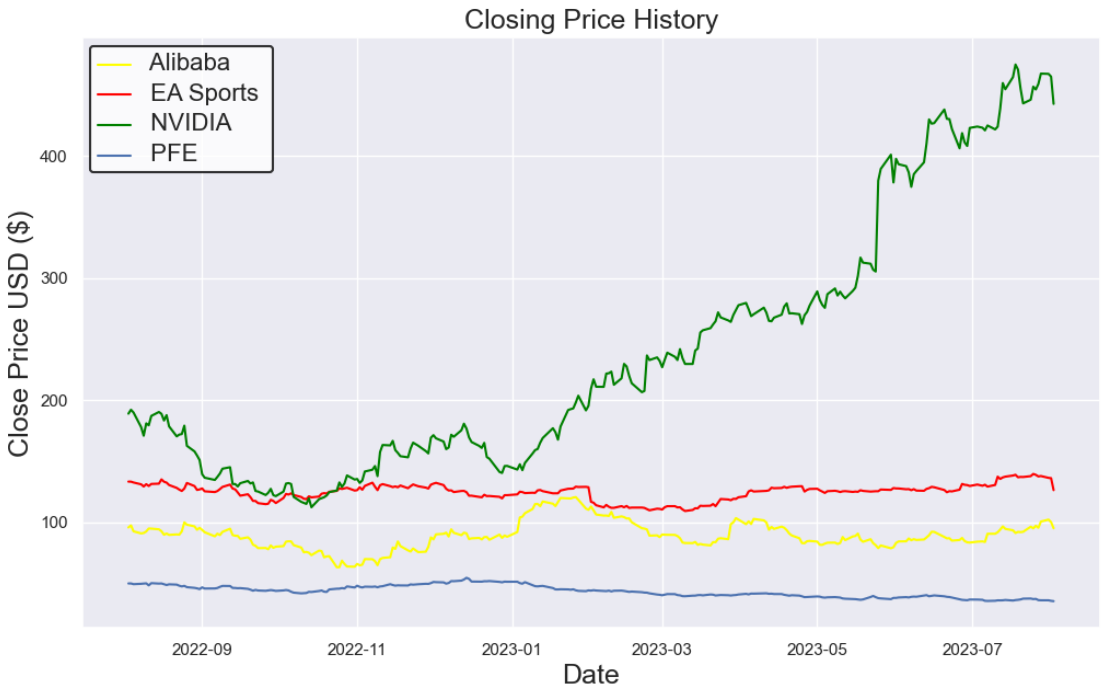

Figure 1: Price Trends of the selected assets.

Among the selected companies, NVIDIA Corporation emerged as the frontrunner, displaying the highest stock price throughout the observed period. Its stock price soared to a remarkable peak of 474.94, demonstrating robust growth and market performance. Conversely, PFE exhibited the lowest level of stock price volatility, characterized by a relatively stable and consistent value, with a minimum recorded at 35.35. Meanwhile, EA and Alibaba witnessed stock price fluctuations within the intermediate range, oscillating between 50 and 150. These observations indicate that the stock prices of EA and Alibaba experienced moderate shifts, reflecting a moderate level of market dynamics and investor sentiment.

2.2. Methodology

Arima is a widely used forecasting model for stock prediction. It effectively captures the autoregressive (AR) and moving average (MA) components in time series data, making it suitable for stock prediction. By incorporating these components, Arima captures patterns, trends, and seasonality in stock price data, making it a valuable tool for stock prediction.

The ARIMA model can be decomposed into three components, the formula form would be $\mathrm{Arima}(p, d, q) = \mathrm{AR}(p) + \mathrm{I}(d) + \mathrm{MA}(q)$. The autoregressive component (AR) represents the relationship between the current observation and a lagged observation, and component (I) represents the differencing process required to transform the temporal data into a stationary form. The moving average component (MA) represents the correlation between the current observation and the error terms of the lagged observations.

LSTM is a viable predictive model for stock prediction. It effectively captures long-term dependencies and handles sequential data, making it well-suited for this task. The organizational structure of LSTM comprises neural units for storage and three gates. This storage unit is capable of preserving relevant information while discarding irrelevant data from the information set. This capability enables LSTM to capture intricate dynamics and non-linear relationships in stock price data, making it a valuable tool for stock prediction.

DOI: 10.54254/2754-1169/57/20230772

The most crucial concepts of LSTM are the three gates, respectively named input gate, forget gate and the output gate. The input gate primarily controls the flow of information that enters the memory cell, and its decision relies on the relevance of the current input information to the previous hidden information. On the other hand, the role within the forget gate is to identify the information that should not be retained and discard it, thereby minimizing interference. As for the output gate, its function is to selectively extract appropriate information from the stored data for the current computation. The formulas for each gate are as follows:

$$I_t = \omega \times (W_i \times [h_{t-1}, X_t]) + \omega b_i \quad (1)$$

$$F_t = \omega \times (W_f \times [h_{t-1}, X_t]) + \omega b_f \quad (2)$$

$$\widetilde{C_t} = \tanh \times (W_c \times [h_{t-1}, X_t]) + \tanh \times b_c \quad (3)$$

$$O_t = \omega \times (W_o \times [h_{t-1}, X_t]) + \omega b_o \quad (4)$$

Where:

I_t, F_t, O_t represents the three different gates mentioned above activation at time step which denotes by t.

W_c, W_i, W_f and W_o are the weight matrices for the input gate connections.

$\widetilde{C_t}$ represents the cell state or the short-term memory while h represents the hidden state or the long-term memory.

X_t represents the input corresponding to time t.

h_{t-1} denotes the hidden state from the previous time step.

b_c, b_i, b_f and b_o are the bias term for the gate.

3. Results

The study first investigated the fitting performance of Arima. After completing data pre-processing and data verification, the time series is differenced twice until it becomes a stationary sequence, obtaining the parameter d for the autoregressive model (AR). Subsequently, parameter q and p are determined through the analysis of the ACF and PACF, respectively. The evaluation of the model is performed using the AIC and the BIC. The corresponding parameters for each dataset are as follows in Table 1:

Table 1: parameters for each dataset.

	Alibaba	EA Sports	NVIDIA	PFE
p	1	1	1	1
d	1	1	1	1
q	1	1	1	1
AIC	(4,3)	(1,0)	(4,3)	(1,0)
BIC	(1,0)	(1,0)	(1,0)	(1,0)

Based on the comprehensive parameter comparison performed on the test dataset, the optimal values for the ARIMA model parameters are determined as (d, p, q) = (1, 1, 1). Subsequently, the ARIMA model is fitted using these optimized parameters, and predictions are made on the training dataset. The predicted values are then compared with the data from the test set, leading to the generation of fitting graphs visualizing the goodness of fit. Furthermore, the Mean Squared Error (MSE) is calculated as a quantitative measure of the prediction accuracy, providing additional insights

DOI: 10.54254/2754-1169/57/20230772

into the performance of the ARIMA model. The outcomes are illustrated in the subsequent Figures 2-5:

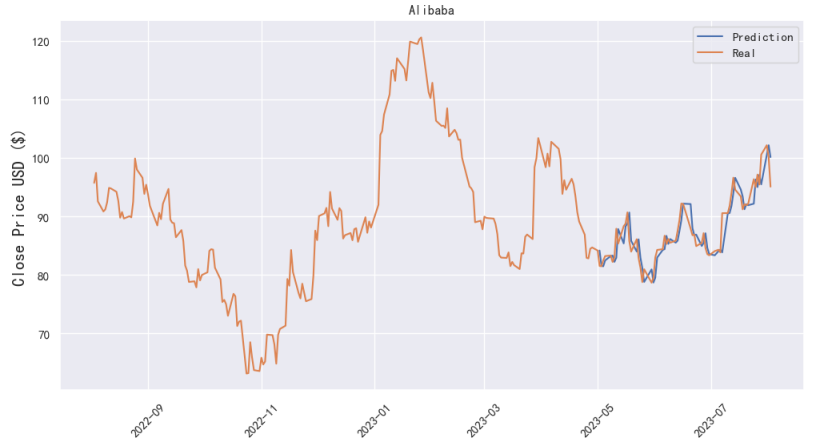

Figure 2: Partial demo of ARIMA performance on Alibaba.

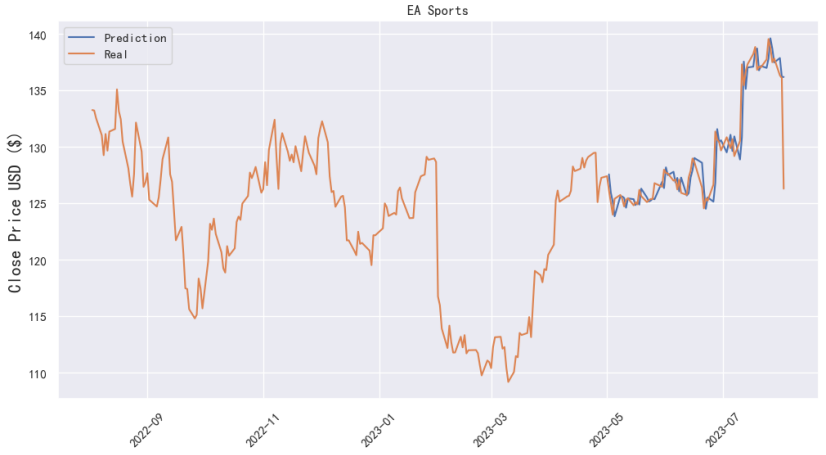

Figure 3: Partial demo of ARIMA performance on EA Sports.

Figure 4: Partial demo of ARIMA performance on NVIDIA.

DOI: 10.54254/2754-1169/57/20230772

Figure 5: Partial demo of ARIMA performance on PFE.

Next, the LSTM model is trained using a sliding window prediction approach in this study. A sliding window of 60 days is chosen as the window size for the prediction model, and it is used to make predictions on the training dataset. After adjusting the model parameters, the final LSTM model consists of one input layer and two LSTM layers. The input layer has a shape of (x_train.shape[1], 1), indicating that it expects input sequences with dimensions (number of time steps, 1). Each LSTM layer has 50 units. The parameter return_sequences=True for the first LSTM layer indicates that it returns sequences as output. The second LSTM layer has return_sequences=False, indicating that it returns a single output. Finally, there are two dense layers, one with 25 units and the other with 1 unit. The model is trained for 20 epochs using Mean Squared Error (MSE) as the assessment indicators. The results are shown in the following Figures 6-9:

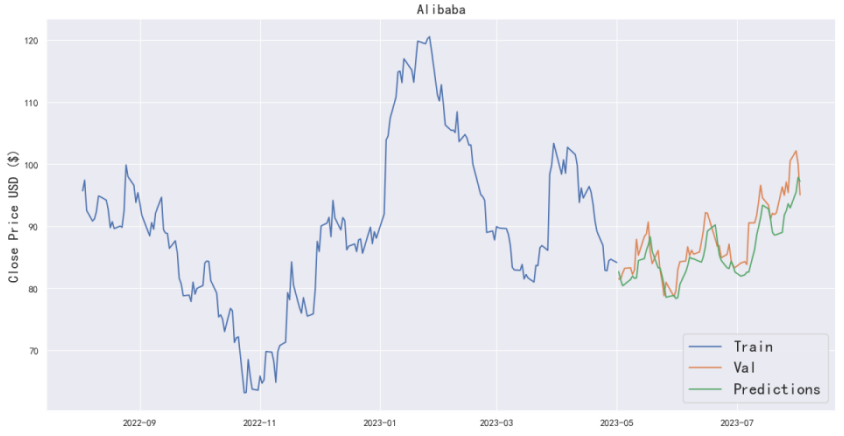

Figure 6: Partial demo of LSTM performance on Alibaba.

DOI: 10.54254/2754-1169/57/20230772

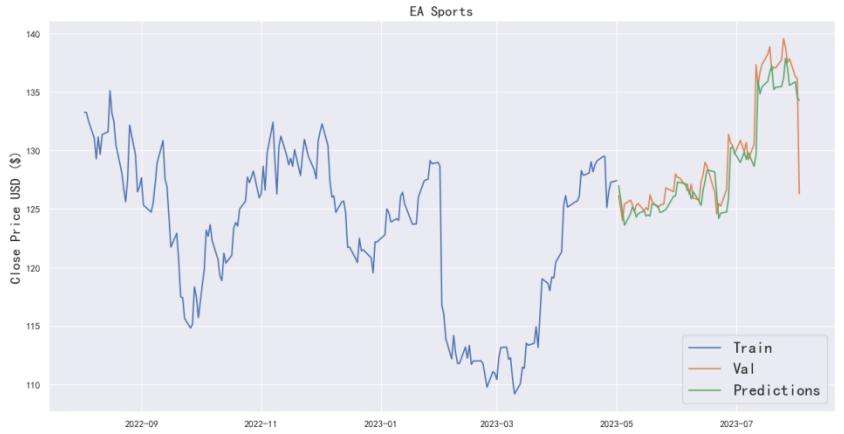

Figure 7: Partial demo of LSTM performance on EA Sports.

Figure 8: Partial demo of LSTM performance on NVIDIA.

Figure 9: Partial demo of LSTM performance on PFE.

Based on these results, such following traits could be observed:

From the displayed fitting graphs, it can be observed that both models exhibit a strong fit to the actual values of the data, following a similar overall trend. However, in comparison, the ARIMA model demonstrates a higher level of conformity to the true values across the four companies. On the other hand, while the LSTM model performs well in fitting the data for EA Sports, its fit is relatively lower for the other three companies (See Table 2).

DOI: 10.54254/2754-1169/57/20230772

Table 2: The Mean Squared Error (MSE) for each model on the respective datasets.

MSE COMPARISON	Alibaba	EA Sports	NVIDIA	PFE
ARIMA	5.319799	3.547143	176.645908	0.296058
LSTM	10.09907	3.992012	492.415274	0.792784

Based on the analysis of the MSE results, it can be tentatively concluded that, under the current parameter settings, the ARIMA model slightly outperforms the LSTM model in the stock prediction of these four companies, indicating a certain advantage of the ARIMA model in this regard. Particularly, the ARIMA model exhibits a greater advantage in the case of Alibaba and NVIDIA, while this advantage is less pronounced for EA and PFE. Both models perform well in the prediction of EA and PFE stocks, but their performance in the case of NVIDIA is relatively mediocre, possibly due to the high volatility of NVIDIA's stock prices. However, regardless of the results, it is recommended to explore alternative prediction models to achieve better fitting results for NVIDIA's stock prediction, as both the LSTM and ARIMA models themselves may not fully meet the requirements.

4. Conclusion

This study examines the closing stock data of four companies representing different industries: Pfizer Inc. (PFE) in the healthcare sector, Alibaba Group Holding Limited in the e-commerce sector, EA Sports in the gaming industry, and NVIDIA Corporation in the AI industry. By constructing ARIMA and LSTM prediction models and comparing them using Mean Squared Error (MSE), it was found that, under the current model parameters, both models perform well in predicting the stock prices of Alibaba, EA, and PFE, while their performance is relatively mediocre for NVIDIA's prediction. Nevertheless, for all four companies, ARIMA outperforms LSTM. Therefore, it is concluded that, among these two models, ARIMA is more suitable for stock price prediction in these four companies. Based on this conclusion, it can be inferred that this approach may also be applicable to other companies within the respective industries.

In conclusion, this study provides a valuable framework and reference for stock price prediction in the healthcare, e-commerce, gaming, and AI industries. Nevertheless, it is crucial to acknowledge that the findings presented in this study are derived from the parameters provided and may not necessarily apply to other parameter settings. Additionally, due to the limited dataset and the number of models employed in this study, further research and data collection are necessary to validate and support these findings. It is also recommended to consider alternative models beyond ARIMA and LSTM for a more comprehensive analysis. Future studies should aim to refine and expand upon these findings to achieve better results.

References

[1] *Upadhyay, A., Bandyopadhyay, G., and Dutta, A. (2012) Forecasting stock performance in indian market using multinomial logistic regression. Journal of Business Studies Quarterly, 3(3), 16.*

[2] *Dutta, A., Bandopadhyay, G., and Sengupta, S. (2012) Prediction of stock performance in the Indian stock market using logistic regression. International Journal of Business and Information, 7(1), 105.*

[3] *Zhong, X., and Enke, D. (2017) Forecasting daily stock market return using dimensionality reduction. Expert systems with applications, 67, 126-139.*

[4] *Hiransha, M., Gopalakrishnan, E. A., Menon, V. K., and Soman, K. P. (2018). NSE stock market prediction using deep-learning models. Procedia computer science, 132, 1351-1362.*

[5] *Adebiyi, A. A., Adewumi, A. O., and Ayo, C. K. (2014) Comparison of ARIMA and artificial neural networks models for stock price prediction. Journal of Applied Mathematics.*

[6] *Jarrett, J. E., and Kyper, E. (2011). ARIMA modeling with intervention to forecast and analyze Chinese stock prices. International Journal of Engineering Business Management, 3, 17.*

DOI: 10.54254/2754-1169/57/20230772

[7] *Ravikumar, S., and Saraf, P. (2020) Prediction of stock prices using machine learning (regression, classification) Algorithms. In 2020 International Conference for Emerging Technology, 1-5.*

[8] *Patel, J., Shah, S., Thakkar, P., and Kotecha, K. (2015) Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert systems with applications, 42(1), 259-268.*

[9] *Lapedes, A., and Farber, R. (1987) Nonlinear signal processing using neural networks: Prediction and system modelling (No. LA-UR-87-2662; CONF-8706130-4).*

[10] *Baştanlar, Y., and Özuysal, M. (2014) Introduction to machine learning. miRNomics: MicroRNA biology and computational analysis, 105-128.*

[11] *Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009) The elements of statistical learning: data mining, inference, and prediction, 2, 1-758. New York: springer.*