International Journal of Informatics Media and Communication Technology
ISSN: Online 2682-2881 & Print 2682-2105

Recurrent Neural Networks with LSTM for Stock Market Index Prediction

Heba M. Ezzat

Faculty of Economics and Political Science, Cairo University, Cairo, Egypt

Correspondence should be addressed to heba.ezzat@feps.edu.eg, h.m.ezzat@gmail.com , ***ORCID ID:*** *0000-0002-0786-8000*

DOI: 10.21608/ijimct.2024.220153.1048

Abstract:

Recurrent neural nets is one of the important deep learning techniques. Efficient Market Hypothesis assumes that financial prices cannot be predicted as they absorb all the news and information. Recently, the enhancement in computational power and learning process by different machine learning techniques facilitates financial market prediction. In this paper, the Long-Short Term Memory (LSTM) is followed to predict financial market index prices and returns. The main index of the Egyptian Stock Exchange; EGX30 is investigated. To the best of our knowledge, no research has been conducted to predict the EGX30 using the LSTM model. Two forecasting models were applied; the Auto-Regressive Integrated Moving Average (ARIMA) and the LSTM for the sake of comparing their performance. The prediction was performed on EGX30 prices and returns. The results show that LSTM performs pretty well in predicting the EGX30 prices and returns. Results illustrate that the LSTM performs better than the ARIMA in both prices and returns.

Keywords

EGX30, financial time series, deep learning, stock returns, ARIMA

1. Introduction

Financial markets play significant role in any economy. Financial markets are very sensitive to any news or rumours. Financial time-series absorb all the news instantly. This hardens predicting future prices for financial assets. This is the main assumption of the efficient market hypothesis (Fama 1970). Recently, many models were suggested to predict future prices of financial markets and they performed well [c.f. (Edwards 2018), (Ezzat 2021), (Kissell 2021), (Shah 2019), (Vaisla 2010), and (Zhong 2017)].

Moreover, there are many characteristics of financial time series such as excess volatility, volatility clustering, heavy tailed-distribution of returns, power-law tails, random-walk prices, and fractal structure that cannot be captured by traditional models (Cont 2001). The new advances in computing powers facilitate predicting financial time-series.

There are many machine learning and data mining techniques that could be followed. Recurrent Neural Networks (RNN) is one of the machine deep learning techniques that aims to mimic human nervous system, such as optical recognition, voice recognition, and pattern recognition (Graves 2013). RNN consists of several layers where the output of each layer could be inserted as the input for the next layer. RNNs principally are composed of three layers; input layer, hidden layer, and output layer. In RNN, the hidden layer is placed between the input and output layers. In the hidden layer(s), the function applies weights to the inputs and transfers them through activation function as output. Activation functions are usually nonlinear, such as sigmoid and tanh, to capture the complex behavior of the data.

In this paper, the Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) will be implemented on the main index of the Egyptian Stock Exchange; EGX30. To the best of our knowledge, no research has been conducted to predict the EGX30 following the LSTM model. The main role of the LSTM model is controlled by a memory cell known as 'cell state' that adapts its state over time [(Hochreiter 1997) and (Olah 2015)]. LSTM model could be used for predicting time series. Some researchers implemented the LSTM for stock market price prediction [c.f. (Althelaya et al. 2018), (Baek 2018), (Liu 2018), and (Nelson 2017)].

Auto-Regressive Integrated Moving Average (ARIMA) model is applied as a benchmark for the performance of the suggested LSTM model. ARIMA is a model that regresses the prices or returns on their own past values, which are its own lags and lagged forecast errors (Hamilton 1994). Thereafter, this model is implemented to predict future prices and returns. There are some weaknesses of ARIMA, such as it is computationally expensive and it has a poor performance for long term prediction. However, ARIMA is one of the most famous traditional models followed for time series forecasting.

The rest of the paper will be divided to four sections. Section 2 is devoted to describe the collected data and their preparation. In section 3, the prediction models are explained. In Section 4, the main results are displayed and discussed. Finally, Section 5 concludes the paper.

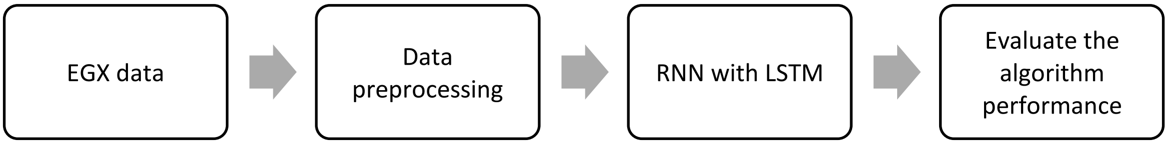
2. Data description and preprocessing

To approach the main purpose, the closing price of the main index EGX30 is collected from investing.com. The data were collected from 5/18/1998 to 5/17/2023 with a total number of

6116 observations. Closing prices and their calculated asset returns are used for prediction.

The frameworks followed to implement the ARIMA and the RNN with LSTM are displayed in Figure 1.

A) ARIMA framework.



B) RNN with LSTM framework.

Figure 1. Framework for the ARIMA and LSTM models.

For data preprocessing, prices are transformed to returns by computing the difference log prices as illustrated by Eq (1).

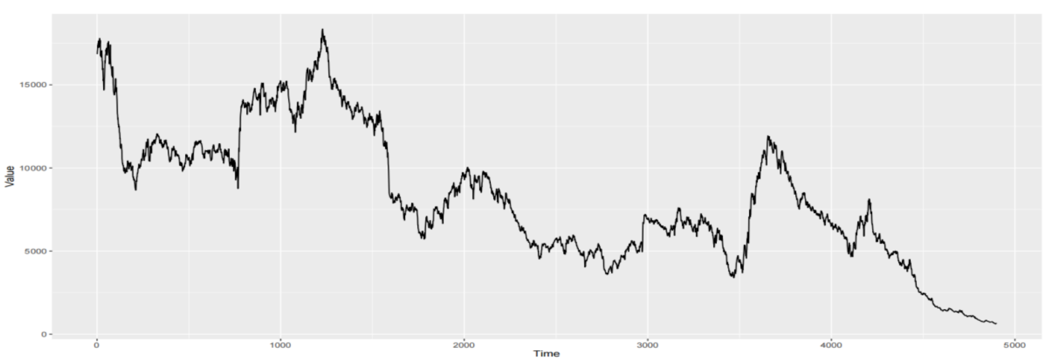
$$r_t = \log(p_t) - \log(p_{t-1}) \quad (1)$$

Both prices and returns are scaled according to Eq (2). Scaling is important to enhance the performance of both models; ARIMA and LSTM.

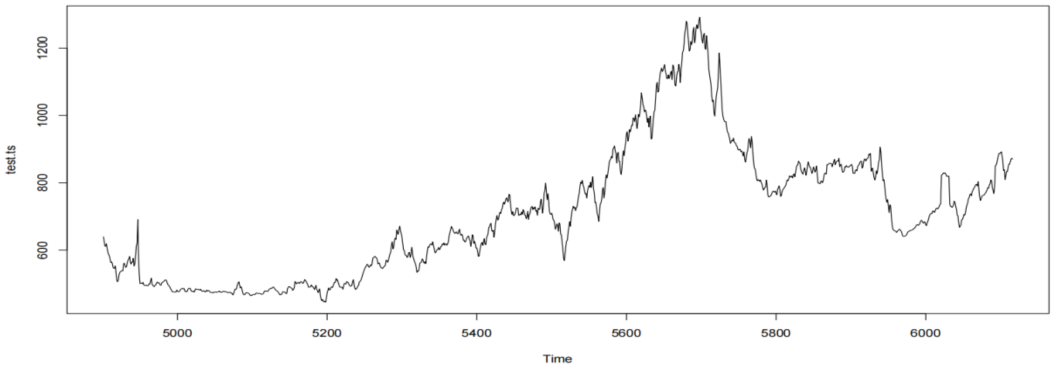
$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2)$$

As the design of this study is to predict future prices and returns, both models ARIMA and LSTM are following the supervised learning. Thereafter, the data were split to 80% for training and

20% for testing and validation. The models are implemented with RStudio software. Figures 2 and 3 show the split of price data to training and test sets before and after scaling.

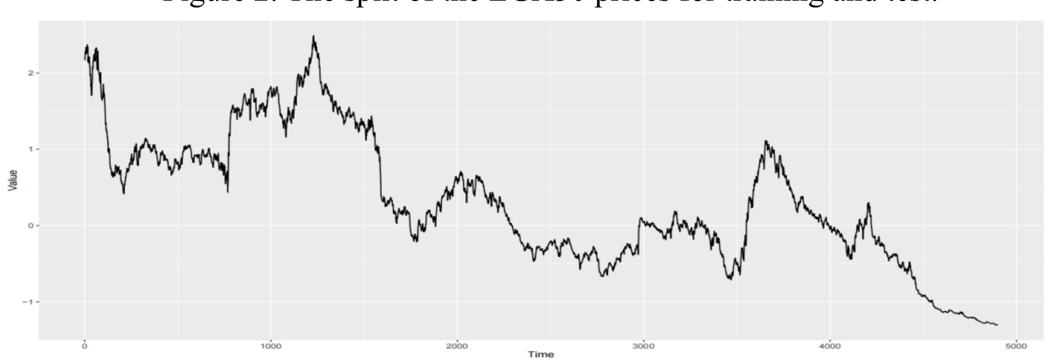


A) The 80% training data of the EGX30

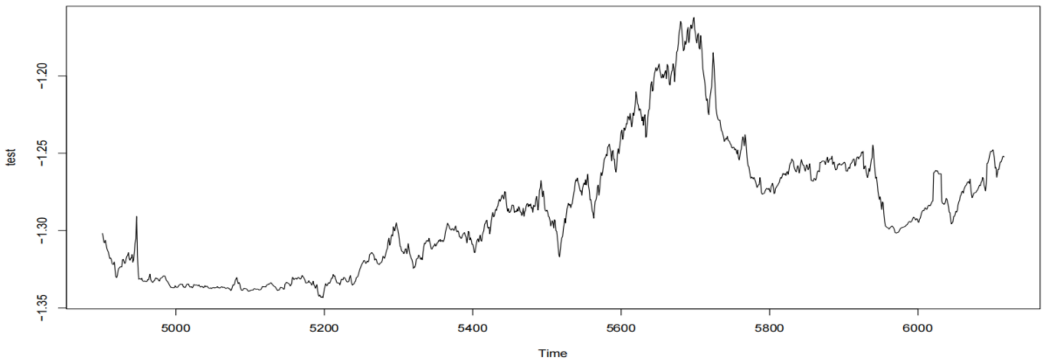


B) The 20% testing data of the EGX30

Figure 2. The split of the EGX30 prices for training and test.



A) The 80% training data of the scaled-price EGX30



B) The 20% testing data of the EGX30

Figure 3. The split of the EGX30 scaled prices for training and test.

3. Prediction models

The main purpose of this study is to implement LSTM model to forecast future prices, returns for the EGX30. To achieve this objective, we need to compare its performance with other forecasting models. Thereby, ARIMA, which is one of the most famous time series forecasting models is applied. The description and explanation of ARIMA and LSTM are presented in the following two subsections.

3.3 ARIMA model

ARIMA is a time series forecasting model composed of two parts (Hamilton 1994);

i. the Auto-Regressive (AR) model as presented in Eq. (3).

Auto-Regressive (AR) model:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \varepsilon \quad (3)$$

Where α is the intercept term and β_i is the coefficient of lag i are estimated by the model. Eq. (3) describes the dependency of Y_t only on its own lags.

ii. Moving Average (MA) model as displayed in Eq. (4).

MA model:

$$Y_t = \alpha + \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2} + \ldots + \phi_q\varepsilon_{t-q} \qquad (4)$$

Eq. (4) represents the dependency of Y_t only on the lagged forecast error. Error terms are the errors of the auto-regressive models of the respective lags.

In ARIMA model, the time series is differenced at least once to transform it to a stationary series. Then, AR part and MA part are integrated to produce the ARIMA model as illustrated by Eq. (5).

ARIMA model:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p}\varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2} + \ldots + \phi_q\varepsilon_{t-q} \qquad (5)$$

3.2 LSTM model

LSTM model is used to identify patterns in data sequences, such as the time series [(Hochreiter 1997) and (Olah 2015)]. Accordingly, LSTM could perform fairly well in predicting stock prices and returns. Deleting the longest retrieved information and substituting it with new data once the memory is full is considered as a real disadvantage of the RNN.

LSTM is considered as a concrete revolution in the field of deep learning, as previous computations are included in the current ones. LSTM, however, solves this problem by filtering selected information in long-term memory, which is stored in the 'cell state'. Moreover, the hidden state stores short-term information from previous calculations.

Figure 4 (https://stackoverflow.com/a/53760729) illustrates the structure of an LSTM unit. LSTM layer is formed by an LSTM cell consisting of several units and many LSTM cells.

Eqs. (6-11) summarize how the LSTM unit's long-term state is computed, its short-term state, and its output at each time step.

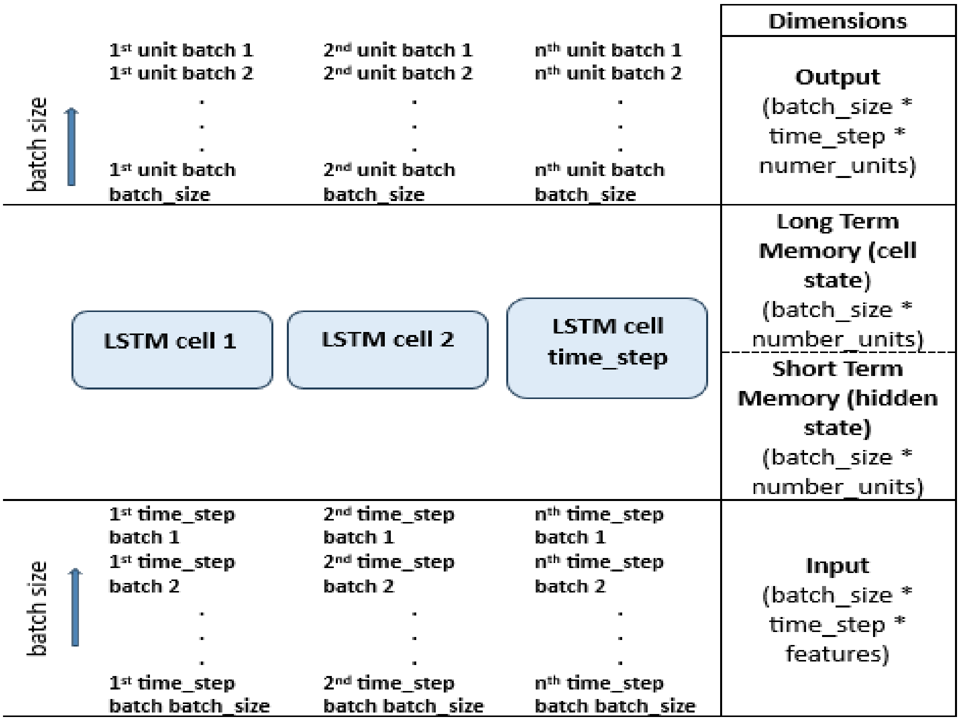


Figure 4. A general LSTM unit.

Input gate: $i_t = \sigma(W_{xi}^T X_t + W_{hi}^T h_{t-1} + b_i)$ (6)

Forget gate: $f_t = \sigma(W_{xf}^T X_t + W_{hf}^T h_{t-1} + b_f)$ (7)

New candidate: $\widetilde{C}_t = \tanh(W_{xc} X_t + W_{hc} h_{t-1} + b_c)$ (8)

Cell state: $C_t = f_t \circ C_{t-1} + i_t \circ \widetilde{C}_t$ (9)

Output gate: $o_t = \sigma(W_{xo} X_t + W_{ho} h_{t-1} + b_o)$ (10)

Hidden gate: $h_t = o_t \circ \tanh(C_t)$ (11)

- W_{xi}, W_{xf}, W_{xc}, W_{xo} are the weight matrices of the three gates, respectively.

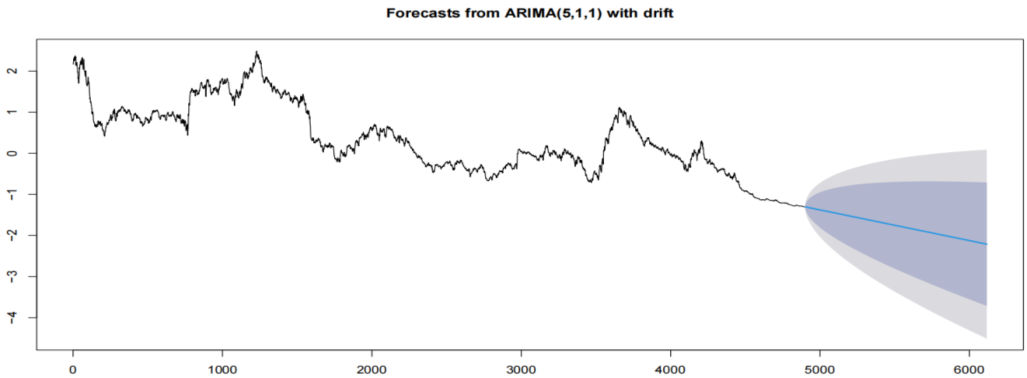
- W_{hi}, W_{hf}, W_{hc}, W_{ho} are the weight matrices of each of the three gates, respectively.
- b_i, b_f, b_c and b_o are the bias terms for each of the three gates, respectively.
- σ is an element-wise sigmoid activation function of the neurons, and tanh is an element-wise hyperbolic tangent activation function of the neurons.
- $\circ$ is the element-wise product.

4. Results and discussion

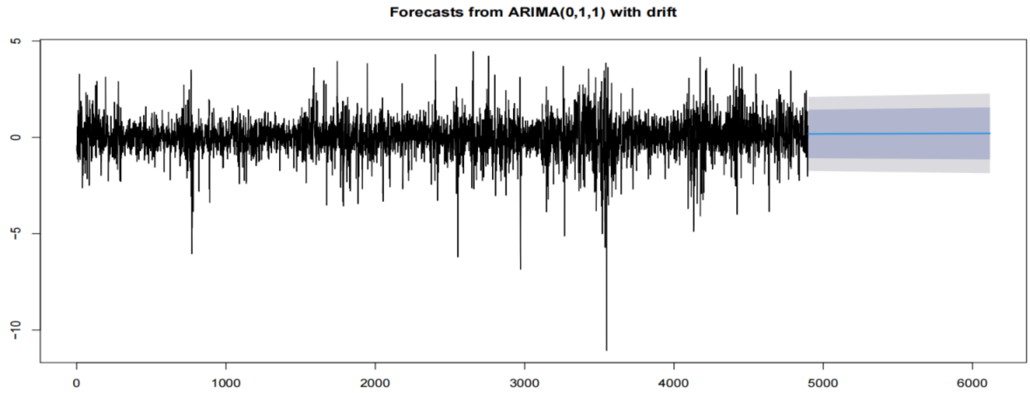
In the following two subsections, the results obtained from implementing the ARIMA and the LSTM models will be presented.

4.1 Results of ARIMA

The results of forecasting prices and returns following ARIMA are displayed in Figure 5.



A) EGX30 scaled prices forecasting.



B) EGX30 scaled returns forecasting.

Figure 5. Price and returns forecasting from ARIMA.

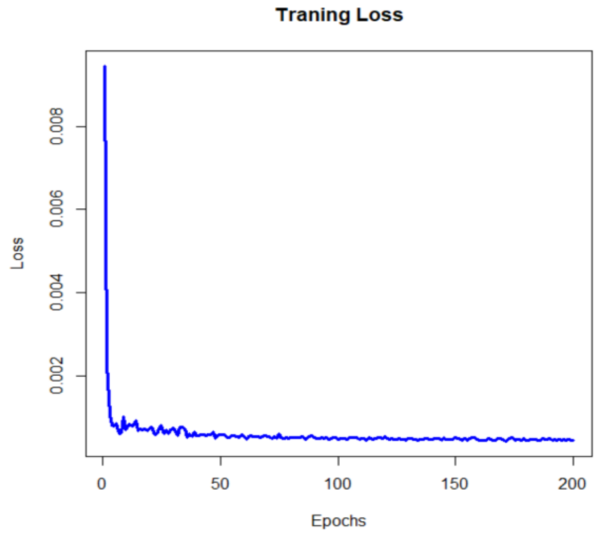
To check the performance of the ARIMA model, Root Mean Square Error (RMSE) is calculated according to the Eq (12). The resulted RMSE are 0.9813 and 0.7708 for prices and returns, respectively.

$$RMSE=\sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \qquad (12)$$

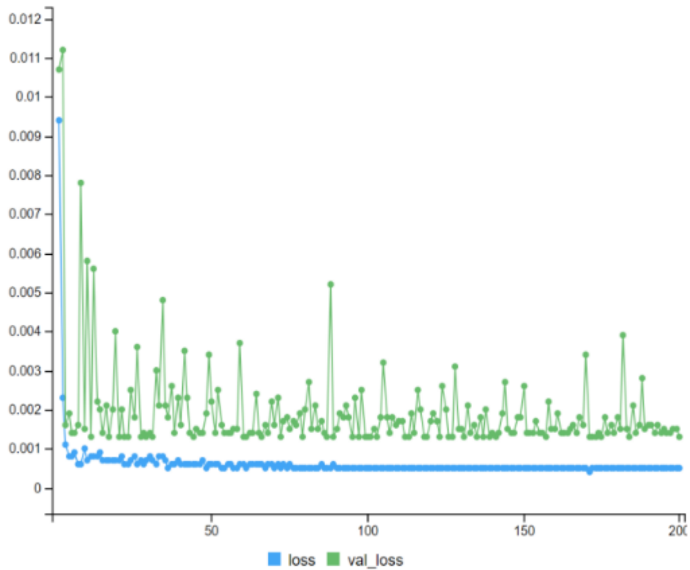
4.2 Results of LSTM

As illustrated above the LSTM is applied on both index prices and returns. To check the extent to how well the LSTM model is working, loss function is computed. Loss function measures how well the machine learning model fits the data set under investigation in terms of predicting the expected values.

In the deep learning, the NN aims at minimizing the loss as it represents the distance between actual values and their predictions. The NNs learn by adjusting weights and biases in a way that decreases loss. Panel A in Figure 6 shows that the loss for training set is decreasing with increasing number of epochs. Panel B in Figure 6 represents loss for training against val_loss for test set. We can observe that both are decreasing with the number of epochs. This indicates that the model is learning and fitting the data fairly well.



A) Loss for the training set



B) Loss for training against loss for test

Figure 6. Training loss and loss against val_loss.

Figure 7 represents the original scaled prices of EGX30 and their train and test prediction by LSTM model. We can notice that the LSTM model is learning very well.

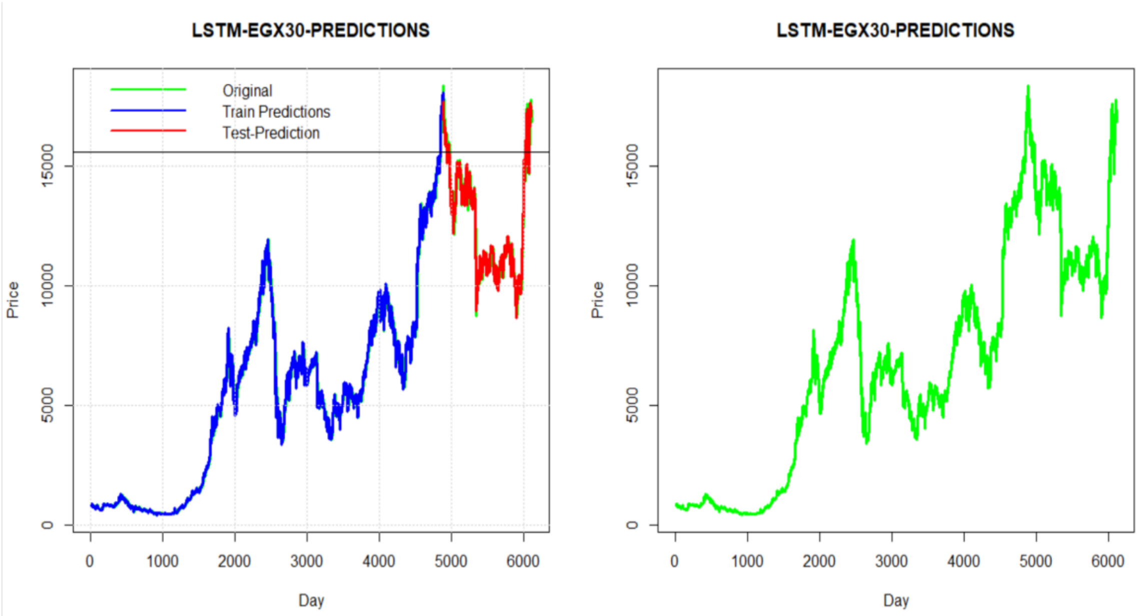
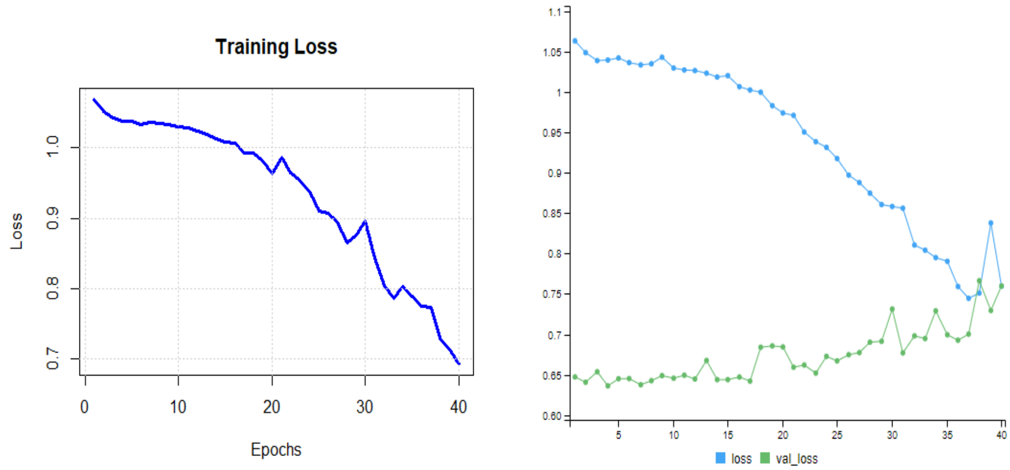


Figure 7. LSTM model for EGX30 prices' predictions.

Figures 8 and 9 display the LSTM results for EGX30 returns. Figure 8 presents the LSTM loss for training in Panel A. The loss function is decreasing with the number of epochs. This indicates that the learning is going well. Also, Panel B shows that the loss for training is decreasing and val_loss is almost stable. This indicates that the LSTM is performing pretty well.

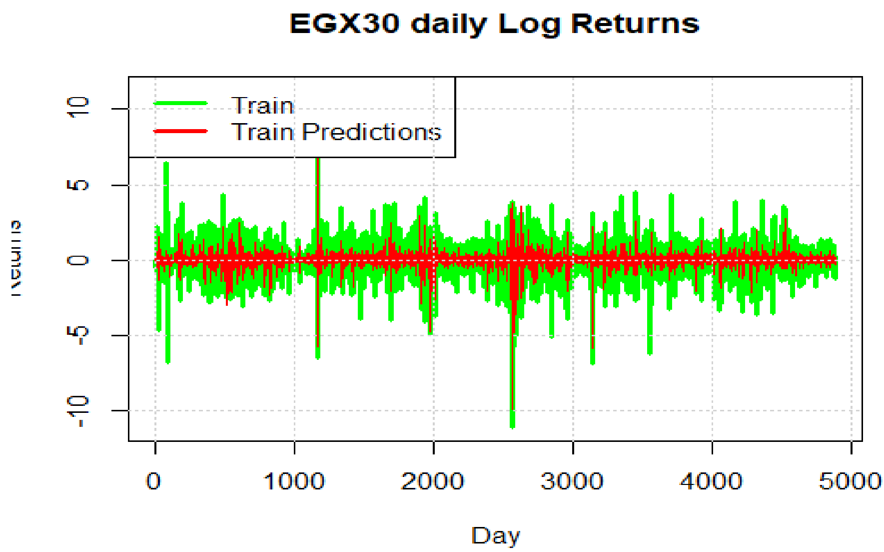


A) Loss for the training set

B) Loss for training against loss for test

Figure 8. Training loss and loss against val_loss.

Figure 9 illustrates the original scaled-returns against train and test predictions. We observe that the model is learning and fitting the data very well.



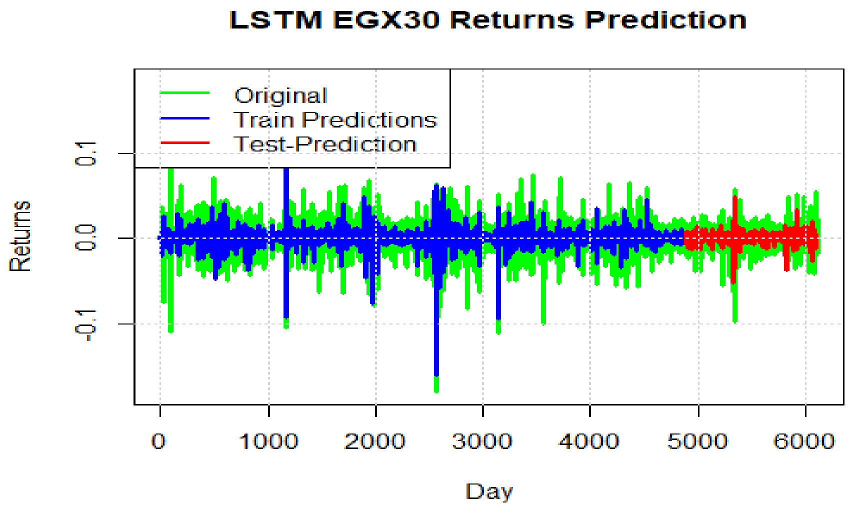


Figure 9. LSTM for EGX30 returns' prediction.

To check the performance of the LSTM model, RMSE is calculated by Eq (12). The resulted RMSE are 0.9438 and 0.7118 for prices and returns, respectively.

5. Conclusion

Predicting financial time series is a very complicated process. This could be due to the complex features of financial time series such random-walk prices, heavy-tailed return distributions, power-law tails, excess volatility, volatility clustering, and fractal structure. Also, financial markets are very sensitive to rumours and news. Financial markets could be affected by factors other than economic status, such as COVID-19 or wars in faraway countries.

However, the advances to predict financial markets are continuing specially with the disruption of deep learning algorithms. RNNs play significant role in deep learning. LSTM model is one of the RNNs with the improvement of the memory status. LSTM can use recent calculations in current learning. In this study, two models for forecasting were applied; the ARIMA and the LSTM. This facilitates the comparison between the performance of both models. The RMSE shows that the LSTM performs better than the ARIMA.

References

Althelaya, K.A., El-Alfy, E.S.M., Mohammed, S.: Evaluation of bidirectional LSTM for short-and long-term stock market prediction. In: 9th International Conference on Information and Communication Systems (ICICS), pp. 151–156. IEEE (2018)

Baek, Y., Kim, H.Y.: ModAugNet: a new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module. Expert Syst. Appl. **113**, 457–480 (2018)

Cont, R., Empirical properties of asset returns: stylized facts and statistical issues.Quantitative Finance 1 (2), 223-236 (2001)

Edwards, R.D., Magee, J., Bassetti, W.C.: Technical Analysis of Stock Trends. CRC Press (2018)

Ezzat, H.M. . Principal Component Regression for Egyptian Stock Market Prediction, The International Journal of Informatics, Media and Communication Technology, 3(1), pp. 23-43. (2021). DOI: 10.21608/ijimct.2021.169612

Fama, E.F.: Efficient capital markets: a review of theory and empirical work. J. Finance **25**(2), 383–417 (1970). http://www.jstor.org/stable/2325486

Graves, A.: Generating sequences with recurrent neural networks. arxiv preprint arxiv: 1308.0850 (2013)

Hamilton, J. . Time Series Analysis. Princeton University Press. (1994). ISBN 9780691042893

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

Kissell, R.: Algorithmic Trading Methods: Applications Using Advanced Statistics, Optimization, and Machine Learning Techniques. Academic Press (2021)

Liu, S., Liao, G., Ding, Y.: Stock transaction prediction modeling and analysis based on LSTM. In: 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 2787–2790. IEEE (2018)

Nelson, D.M., Pereira, A.C., de Oliveira, R.A.: Stock market's price movement prediction with LSTM neural networks. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 1419–1426. IEEE (2017)

Olah, C.: Understanding LSTM networks (2015)

Shah, D., Isah, H., Zulkernine, F.: Stock market analysis: a review and taxonomy of prediction techniques. Int. J. Financ. Stud. **7**(2), 26 (2019)

Vaisla, K.S., Bhatt, A.K.: An analysis of the performance of artificial neural network technique for stock market forecasting. Int. J. Comput. Sci. Eng. 2(6), 2104– 2109 (2010)

Zhong, X., Enke, D.: Forecasting daily stock market return using dimensionality reduction. Expert Syst. Appl. 67, 126–139 (2017)