Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/12S118

A Comparative Study on Market Index Prediction: Long Short-Term Memory (LSTM) vs. Decision Tree Model

Afrah Al-Bossly[1], *Manahill I. A. Anja*[2], *Abdelgalal O. I. Abaker*[3,*], *Hago E. M. Ali*[4], *and Salem Alkhalaf*[5]

[1]Department of Mathematics, College of Science and Humanities in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj, 11942, Saudi Arabia
[2]Computer Sciences Program, Department of Mathematics, Turabah University College, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia
[3]Applied College, Khamis Mushait, King Khalid University, Abha, Saudi Arabia
[4]Department of Business Administration, Faculty of Science and Humanity Studies. Sulail, Prince Sattam bin Adalaziz University, Saudi Arabia
[5]Department of Computer, College of Science and Arts in Ar Rass, Qassim University, Ar Rass, Saudi Arabia

Received: 17 Sep. 2023, Revised: 20 Oct. 2023, Accepted: 7 Nov. 2023.
Published online: 1 Dec. 2023.

Abstract: The main objective of this article is to develop a linear exponential function risks in Saudi banks (LINEXLF) to estimate the shape parameter, reliability, and hazard rate functions of the Pareto distribution based on Type II Censored Data. By weighting LINEX loss function to produce a modified loss function called weighted linear exponential (WLINEXLF) loss function. We then use WLINEXLF to derive the shape parameter, reliability, and hazard rate functions of the Pareto distribution. Furthermore, to examine the performance of the proposed method WLINEXLF we conduct a Monte Carlo simulation. The comparison is between the proposed method and other methods including maximum likelihood estimation (MLE) and Bayesian estimation under the squared error loss function. The results of the simulation show that the proposed method WLINEXLF in this article has the best performance in estimating shape parameter, reliability, and hazard rate functions, according to the smallest values of mean squared error (MSE). This result means that the proposed method can be applied in real data in banking industrial sectors. This paper aims to use the modified loss function to estimate the shape parameter, reliability $R(t)$, and hazard rate functions $h(t)$ in Saudi banks of the Pareto distribution based on Type II Censored Data.

Keywords: Bayes, Pareto Distribution, Type II Censored Data, Weighted LINEX, Prediction, Estimation, Decision Trees.

1 Introduction

Anticipating the future dynamics of financial markets has perpetually been an intriguing and elusive challenge. The emergence of machine learning algorithms in recent years has provided novel resources for addressing this difficulty. Notably, Long Short-Term Memory (LSTM) networks and decision tree models are prominent and commonly employed methods. Machine learning models, equipped with advanced algorithms and the ability to detect complex patterns, significantly enhance the precision of market trend predictions [1]. These models can analyze large datasets, identify subtle trends, and adapt to changing market conditions, thus providing a more advanced understanding of expected future changes. Predicting variations in the stock market is a complex undertaking, and numerous machine-learning methods can be utilized for this purpose [2]. LSTM and Decision Trees are two separate algorithms that can be used for predicting stock market trends [3]. Each method has its unique strengths and weaknesses. This work aims to employ LSTM and Decision Trees to forecast stock market movements. The study conducted by M. Nabipour and colleagues [4] seeks to identify the optimal

[*]Corresponding author e-mail: aoadrees@kku.edu.sa

machine-learning method for predicting stock market trends with maximum efficiency. They conduct a comprehensive analysis of different classifiers using multiple stock market datasets. They specifically concentrate on assessing data from four prominent stock market indices, including NASDAQ, NYSE, Nikkei, and FTSE. The study employs data from these prominent exchanges to evaluate the efficacy of several machine-learning techniques in forecasting stock market trends.

Since March 24, 2010, Yahoo Finance has served as the main provider of daily stock data for a period of ten years. The study conducted by D. Radojičić and N. Radojičić [5] examines a wide variety of algorithms. This encompasses supervised and unsupervised machine learning methods, ensemble algorithms, time series analysis techniques, and deep learning algorithms. The primary aim of their research is to forecast stock values and tackle classification difficulties. The following are some of the most important contributions made by this review study:

1. The investigation of implementing machine learning and deep learning models within the banking sector.
2. The provision of a complete framework for predicting and categorizing stock prices, employing an ensemble model called "Random Forest + XG-Boost + LSTM" to forecast TAINIWALCHM and AGROPHOS stock prices.
3. A comparative analysis with other well-known machine learning and deep learning models.

In a study conducted by D. Radojičić, N. Radojičić, and T. Rheinländer [6], the usefulness of recurrent neural networks in creating a categorization model for the stock market was examined. Their research employed data that accurately replicates the intricate limit order book of the Nasdaq stock market. Following the extraction of order book attributes from the original dataset, they utilized feature selection techniques including stochastic universal sampling and conditional entropy. The objective of these strategies was to identify characteristics that could provide valuable insights for the purpose of data classification.

The research conducted by C. A. Hargreaves and C. Leran [7] use the LSTM technique to forecast changes in stock prices in the Australian Stock Market and to determine stocks that are most suitable for a lucrative investment portfolio. Their methodology entailed scrutinizing 400 distinct stocks and handpicking the foremost five for possible investing and trading purposes. They employed predictions from LSTM, Regression Tree (CART), and Auto-Regressive Integrated Moving Average (ARIMA) models to inform their stock selections. The findings demonstrated that LSTM, a sophisticated deep learning methodology, outperformed both CART and ARIMA-Time Series algorithms in terms of performance. The LSTM model demonstrated its effectiveness by a 35% return rate, a Sharpe ratio of 2.13, and a maximum drawdown of 0.34. Importantly, these metrics surpassed the return rate of 21% observed in the Australian market index. LSTM networks demonstrated greater accuracy in forecasting stock values compared to ARIMA time series and CART regression trees, possibly because to their enhanced ability to process sequential data, enabling effective discrimination between vital and non-essential information. In addition, the LSTM model exhibited higher return stability in comparison to the ARIMA model. The capacity to identify non-linear patterns in data provides it with a clear edge over conventional models in predicting stock portfolios. The LSTM model's superior capabilities allow it to continuously exceed the stock market index and provide sustained returns across three distinct time periods, making it more effective than prior models. Nikou and et. al, want to evaluate the precision of machine learning algorithms in stock market applications. Their research [8] use a dataset that includes the monthly closing prices of the iShares MSCI United Kingdom exchange-traded fund. The dataset spans from January 2015 to June 2018. The study entails the utilization of four separate machine-learning algorithm models for the goal of prediction. The study's findings demonstrate that the deep learning methodology surpasses other methods in terms of predicted accuracy. The support vector regression method exhibits a decreased error rate compared to both the neural network and random forest approaches, positioning it as the second most efficacious approach in this particular setting.

In the study of Yang and Zhai [9] propose a sophisticated deep learning model that accurately forecasts the direction of price movements by analyzing historical financial time series data. This method utilizes a convolutional neural network (CNN) to extract features and a LSTM network to make predictions. The approach employs a three-dimensional CNN to process input data, including time series data. Processing of technical indicators and analyzing the correlation between stock indexes are components of the methodology. The technical indicators within a three-dimensional input tensor are transformed into deterministic trend signals during the transformation process. After that, the stock indexes are ordered according to the Pearson product-moment correlation coefficient (PPMCC). A fully connected network is applied to train the CNN during the training phase, resulting in the generation of a feature vector. This vector is used as input for the LSTM that is concatenated. The findings of the experiments demonstrate that the framework significantly outperforms other models considered state-of-the-art in predicting the direction in which stock prices will move.

The structure of this paper is as follows: The first portion contains the introduction, the second section includes the literature review, and sections three and four consist of the clarification of the key principles of LSTM and the decision tree, respectively. Section five displays the accuracy measurement. Section six involves the analysis of data, and the conclusion is the final segment.

2 Long Short-Term Memories

LSTM is a specific sort of recurrent neural network (RNN) structure that addresses the issue of the vanishing gradient problem commonly found in conventional RNNs [10]. LSTMs excel at capturing extended relationships in sequential data, rendering them highly suitable for tasks such as time series prediction, natural language processing, and, in our specific scenario, predicting stock market patterns [11]. Below is a concise elucidation of the fundamental principles that underlie LSTM:

Memory Cells: LSTM networks are equipped with specialized units known as memory cells, capable of retaining information for extended durations [12]. These cells possess an autonomous memory system capable of storing and retrieving information as required.

Gates: LSTM models consist of three distinct gates that regulate the information flow: the input gate, the forget gate, and the output gate [13].

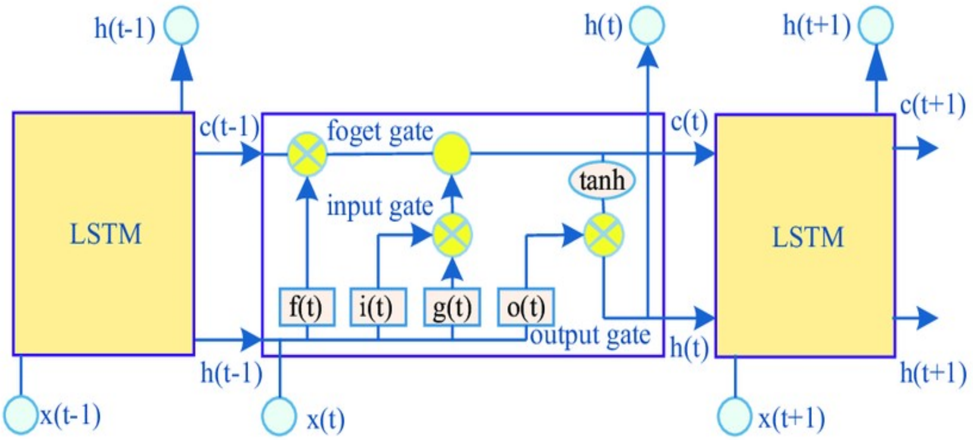


Fig. 1. Long short-term memory (LSTM) cell architecture [14].

3 Decision Tree

A Decision Tree is a type of method in machine learning that is used for supervised tasks, specifically for classifying and predicting values [15]. The fundamental principles of a Decision Tree entail iteratively dividing the data into subsets according to the values of the features, resulting in a hierarchical structure where each internal node signifies a decision based on a feature, each branch signifies an outcome of that decision, and each leaf node signifies the ultimate prediction or classification [16].

Decision Tree Structure: The Decision Tree is composed of nodes, with each node representing a decision or a test on a certain property [17]. The connection between nodes is established through branches, forming a hierarchical tree structure. The topmost node is referred to as the root node, while the nodes at the bottom are known as leaf nodes.

Decision Nodes (Internal Nodes): Every internal node in the tree corresponds to a decision made based on the value of a certain property [18]. The decision is usually a dichotomous selection, such as "Does feature X exceed a certain threshold?" Branches serve as connectors between decision nodes, symbolizing the potential results of the decision.

A branch connects to either a further decision node or a terminal leaf node.

Branches serve as connectors between decision nodes, symbolizing the potential results of the decision.

A branch connects to either a further decision node or a terminal leaf node [19].

Terminal Nodes: Terminal nodes, also known as leaf nodes, serve as the ultimate output or prediction in a tree structure. In classification tasks, every leaf node corresponds to a certain class label.

In regression tasks, the leaf nodes store the forecasted numerical values [20].

Splitting Criteria: The algorithm determines the optimal characteristic to divide the data at every decision node. The objective of the splitting criteria is to optimize the uniformity within each subset of the data, guaranteeing that samples within the same subset exhibit greater similarity in relation to the target variable [21].

Information Gain and Reduction in Variance for regression: The Decision Tree evaluates the effectiveness of different splits using measures such as Information Gain for classification tasks and Reduction in Variance for regression tasks [22]. Information Gain quantifies the reduction in uncertainty regarding the target variable following a split.

Pruning (Optional): Choice Trees are susceptible to overfitting, which means they can incorporate irrelevant details from the training data. Pruning is the process of eliminating branches or nodes from a tree in order to enhance its ability to generalize to unfamiliar input [23].

Interpretability: Decision Trees possess a high degree of interpretability due to the clear representation of decision logic in the tree structure. Users can readily comprehend and interpret the process of decision-making [24].

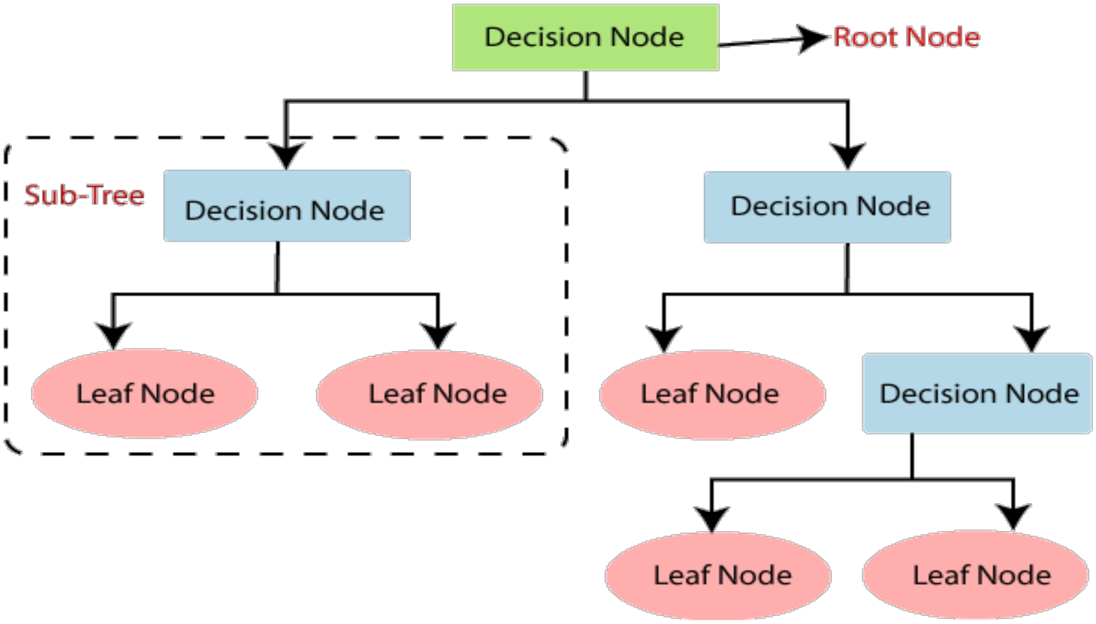


Fig. 2. Decision Tree Structure.

4 The accuracy measurement

We utilized sMAPE, MASE, MAPE in order to evaluate the accuracy of the forecast. Following are several formulas that can be utilized in order to ascertain sMAPE, MASE, and MAPE. [25].

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i|}{|y_i| + |\hat{y}_i|}, \tag{1}$$

$$\mathrm{MASE} = \frac{\frac{1}{n}\sum_{i=1}^{n}|e_i|}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|}, \tag{2}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| . \tag{3}$$

5 Data analysis

In order to forecast stock market performance in the Kingdom of Saudi Arabia, we utilize historical data spanning from May 1, 2018 to September 13, 2023. The information was extracted from the digital Tadawal webpage of the Kingdom of Saudi Arabia**.** Our analysis involves the implementation of a decision tree and the LSTM model.

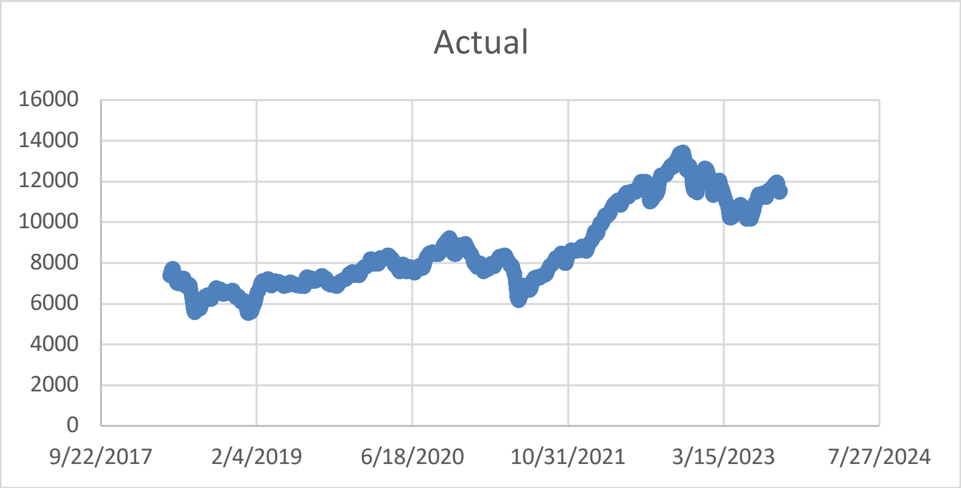
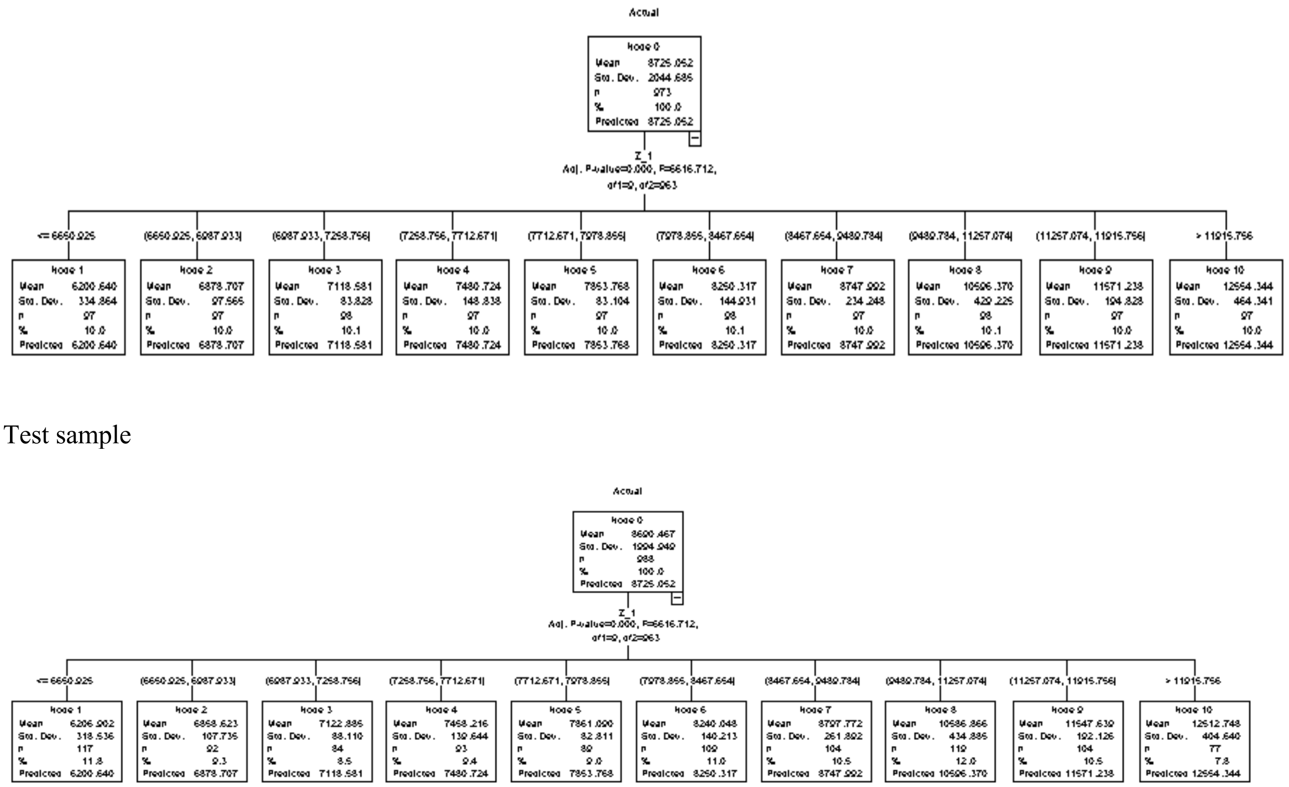


Fig. 3. Temporal progression of the stock market.

Table 1.Response Information.

Mean	St. Dev.	Minimum	Q1	Median	Q3	Maximum
8709.06	2019.83	5540.45	7110.52	8025.06	10576.4	13440.7

Training Sample



Test sample

Fig. 4: Train and testing sample for Decision Tree

Both the training sample and the testing sample were included in the created decision tree.

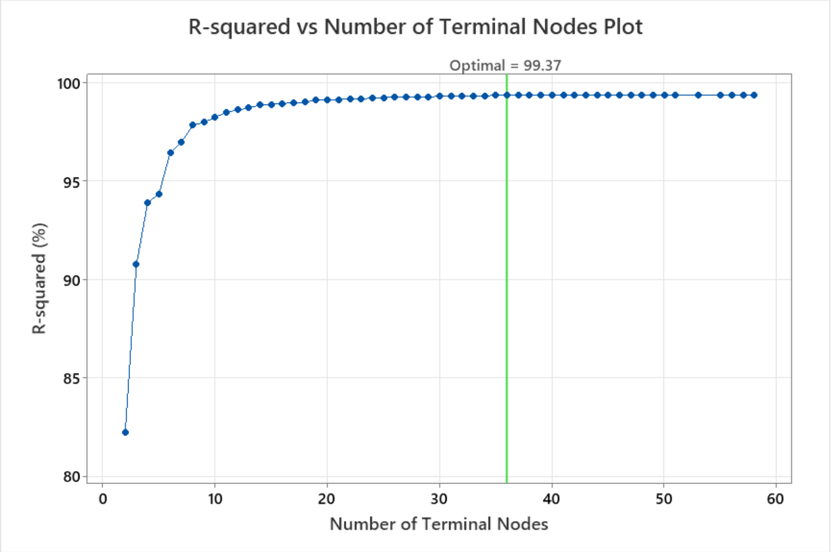


Fig. 5. R-square vs. number of terminal nodes Plot.

Table 2. Model Summary.

Statistics	Training	Test
R-squared	99.50%	99.37%
Root mean squared error (RMSE)	142.5249	159.9368
Mean squared error (MSE)	20313.3342	25579.7956
Mean absolute deviation (MAD)	103.4202	116.4316
Mean absolute percent error (MAPE)	0.0122	0.0136

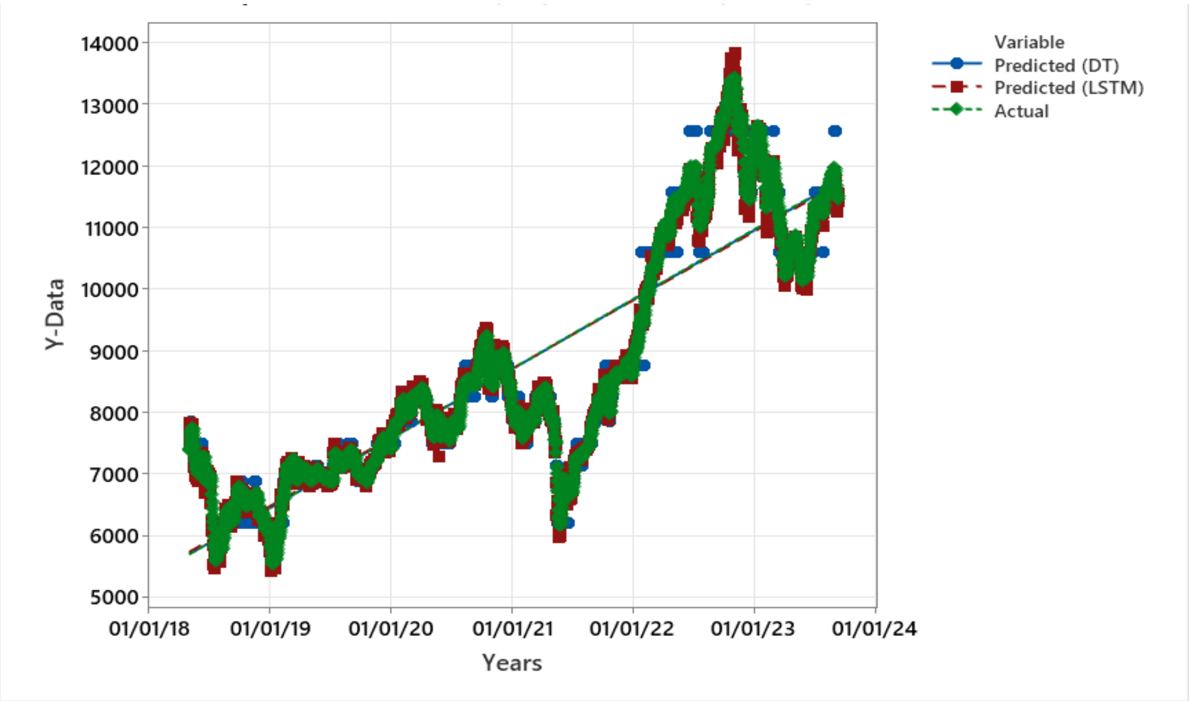


Fig. 6. The scatterplot of predicted DT, LSTM and Actual vs. Years.

6 Conclusions

This thorough investigation evaluated the effectiveness of Long Short-Term Memory (LSTM) and Decision Tree models in predicting market trends, using historical stock market data from the Kingdom of Saudi Arabia spanning from May 1, 2018, to September 13, 2023. The study aimed to assess the prediction efficacy, reliability, and feasibility of these models, employing key performance metrics such as accuracy, (MASE), (MAPE), and (SMAPE) to gauge their effectiveness in predicting market index movements. The findings highlight crucial observations regarding the relative precision of the two models and their effectiveness in diverse market circumstances. Specifically, the Decision Tree model demonstrated superior accuracy in predicting stock market movements in the Kingdom of Saudi Arabia, evident in its lower MAPE of 0.01393281 compared to the LSTM model's MAPE of 0.021172901. Both models, as depicted in Figure 6, showed a high degree of proximity to actual values, supported by extensive evidence demonstrating their superior accuracy in forecasting stock market movements within the specified timeframe in the Kingdom of Saudi Arabia. Consequently, both models yield more precise predictions.

Acknowledgement:

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number RGP2/462/44

References

[1] M. M Taye, Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers., **12(5)**, 91(2023).

[2] E. K., Ampomah, G.,Nyame, Z., Qin, P. C.,Addo, E. O.,Gyamfi, & M. Gyan, Stock market prediction with gaussian naïve bayes machine learning algorithm. Informatica., **45(2)**, (2021).

[3] R. Zhang, LSTM-based Stock Prediction Modeling and Analysis. In 2022 7th International Conference on Financial Innovation and Economic Development (pp. 2537-2542). (2022, March).

[4] M., Nabipour, P., Nayyeri, H., Jabani, S.,Shahab, & A., Mosavi, Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. IEEE Access., **8**, 150199-150212 (2020).

[5] G., Sonkavde, D. S., Dharrao, A. M., Bongale, S. T., Deokate, D., Doreswamy, & S. K Bhat, Forecasting stock market prices using machine learning and deep learning models: a systematic review, performance analysis and discussion of implications. International Journal of Financial Studies., **11(3)**, 94 (2023).

[6] D., Radojičić, N., Radojičić, & T. Rheinländer, A comparative study of the neural network models for the stock market data classification—A multicriteria optimization approach. Expert Systems with Applications., **238**, 122287 (2024).

[7] C. A., Hargreaves, & C. Leran, Stock Prediction Using Deep Learning with Long-Short-Term-Memory Networks. Int. J. Electron. Eng. Comput. Sci., **5(3)**, 22-32 (2020).

[8] M. Nikou, G., Mansourfar, & J. Bagherzadeh, Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. Intelligent Systems in Accounting, Finance and Management., **26(4)**, 164-174 (2019).

[9] C.Yang, J., Zhai, & G. Tao, Deep learning for price movement prediction using convolutional neural network and long short-term memory. Mathematical Problems in Engineering., (2020).

[10] H. Fei, & F. Tan, Bidirectional grid long short-term memory (bigridlstm): A method to address context-sensitivity and vanishing gradient. Algorithms., **11(11)**, 172.2(2018).

[11] S., Sun, Y., Wei, & S. Wang, AdaBoost-LSTM Ensemble Learning for Financial Time Series Forecasting. Computational Science – ICCS 2018, 590–597. doi:10.1007/978-3-319-93713-7_55. (2018).

[12] A. S., PODDA, & D. R., RECUPERO, Explainable Machine Learning Exploiting News and Domain-Specific Lexicon for Stock Market Forecasting.

[13] X. H., Le, H. V., Ho, G., Lee, & v., Jung, Application of long short-term memory (LSTM) neural network for flood forecasting. Water., **11(7)**, 1387 (2019).

[14] Z., Kong, Y., Cui, Z., Xia, & H. Lv, Convolution and long short-term memory hybrid deep neural networks for remaining

useful life prognostics. Applied Sciences., **9(19)**, 4156 (2019).

[15] B., Charbuty, & A. Abdulazeez, Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends., **2(01)**, 20-28 (2021).

[16] W., Chang, Y., Liu, Y., Xiao, X., Yuan, X., Xu, S., Zhang, & S. A., Zhou, machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics., **9(4)**, 178 (2019).

[17] L.; Benos, A.C.; Tagarakis, G., Dolias, R., Berruto, D., Kateris, D., Bochtis., Machine Learning in Agriculture: A Comprehensive Updated Review. Sensors., **21**, 3758 (2021).

[18] M., Christofi, V., Pereira, S., Tarba, A., Makrides, & E. Trichina., Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. The International Journal of Human Resource Management., **33(6)**, 1237-1266 (2022).

[19] W. M.Tsai, H., Zhang, , E., Buta, S., O'Malley, & R. A., Gueorguieva, modified classification tree method for personalized medicine decisions. Statistics and its Interface., **9(2)**, 239 (2016).

[20] M., Tayefi, H., Esmaeili, M. S., M. S., Karimian, A. A., Zadeh, M., Ebrahimi, M., Safarian, & M. Ghayour-Mobarhan., The application of a decision tree to establish the parameters associated with hypertension. Computer methods and programs in biomedicine., **139**, 83-91 (2017).

[21] M., Moon, & S. K. Lee, applying of decision tree analysis to risk factors associated with pressure ulcers in long-term care facilities. Healthcare informatics research., **23(1)**, 43-52 (2017).

[22] C. C., Chern, Y. J., Chen, & B. Hsiao, Decision tree–based classifier in providing telehealth service. BMC medical informatics and decision making., **19**, 1-15 (2019).

[23] A., Gheondea- Eladi, Patient decision aids: a content analysis based on a decision tree structure. BMC medical informatics and decision making., **19**, 1-15 (2019).

[24] P. U., Kasbekar, P., Goel, & S. P., Jadhav, A decision tree analysis of diabetic foot amputation risk in Indian patients. Frontiers in endocrinology., **8**, 25 (2017)

[25] H., Mandelkow, J. A., De Zwart, & J. H., Duyn, Linear discriminant analysis achieves high classification accuracy for the BOLD fMRI response to naturalistic movie stimuli. Frontiers in human neuroscience., **10**, 128 (2016).