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

Afran Al-Bossly
Department of Mathematics, College of Science and Humanities in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj, 11942, Saudi Arabia, aoadrees@kku.edu.sa

Manahill I. A. Anja
Computer Sciences Program, Department of Mathematics, Turabah University College, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia, aoadrees@kku.edu.sa

Abdelgalal O. I. Abaker
Applied College, Khamis Mushait, King Khalid University, Abha, Saudi Arabia, aoadrees@kku.edu.sa

Hago E. M. Ali
Department of Business Administration, Faculty of Science and Humanity Studies. Sulail, Prince Sattam bin Adalaziz University, Saudi Arabia, aoadrees@kku.edu.sa

Salem Alkhalaf
Department of Computer, College of Science and Arts in Ar Rass, Qassim University, Ar Rass, Saudi Arabia, aoadrees@kku.edu.sa

Follow this and additional works at: https://digitalcommons.aaru.edu.jo/jsap

Recommended Citation
DOI: http://dx.doi.org/10.18576/jsap/12S118
Available at: https://digitalcommons.aaru.edu.jo/jsap/vol12/iss4/18

This Article is brought to you for free and open access by Arab Journals Platform. It has been accepted for inclusion in Journal of Statistics Applications & Probability by an authorized editor. The journal is hosted on Digital Commons, an Elsevier platform. For more information, please contact rakan@aaru.edu.jo, marah@aaru.edu.jo, u.murad@aaru.edu.jo.
A Comparative Study on Market Index Prediction: Long Short-Term Memory (LSTM) vs. Decision Tree Model

Afrah Al-Bossly¹, Manahill I. A. Anja², Abdelgalal O. I. Abaker³*, Hago E. M. Ali⁴, and Salem Alkhalaf⁵

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

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.

**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].

![Decision Tree Structure](image-url)

**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].

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

\[
MASE = \left( \frac{1}{n} \sum_{i=1}^{n} |e_i| \right) \left( \frac{1}{n-1} \sum_{i=2}^{n} |y_i - y_{i-1}| \right)^{-1},
\]

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{e_i}{|y_i|}.
\]

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.

<table>
<thead>
<tr>
<th>Mean</th>
<th>St. Dev.</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>8709.06</td>
<td>2019.83</td>
<td>5540.45</td>
<td>7110.52</td>
<td>8025.06</td>
<td>10576.4</td>
<td>13440.7</td>
</tr>
</tbody>
</table>

Training Sample

![Training Sample Image]

Test sample

![Test Sample Image]

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.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>99.50%</td>
<td>99.37%</td>
</tr>
<tr>
<td>Root mean squared error (RMSE)</td>
<td>142.5249</td>
<td>159.9368</td>
</tr>
<tr>
<td>Mean squared error (MSE)</td>
<td>20313.3342</td>
<td>25579.7956</td>
</tr>
<tr>
<td>Mean absolute deviation (MAD)</td>
<td>103.4202</td>
<td>116.4316</td>
</tr>
<tr>
<td>Mean absolute percent error (MAPE)</td>
<td>0.0122</td>
<td>0.0136</td>
</tr>
</tbody>
</table>

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


