Fırat Üniversitesi Müh. Bil. Dergisi
36(1), 25-34, 2024

Araştırma Makalesi
https://doi.org/10.35234/fumbd.1357613

Stock Price Forecasting Using Machine Learning and Deep Learning Algorithms: A Case Study for the Aviation Industry

Yunus Emre GÜR[1*]
[1] Yönetim Bilişim Sistemleri Bölümü, İktisadi ve İdari Bilimler Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye
[*1] yegur@firat.edu.tr

(Geliş/Received: 09/09/2023; Kabul/Accepted: 10/10/2023)

Abstract: With technological advances, humans are constantly generating data through various electronic devices and sensors, and this data is stored in digital environments. A vast amount of data has served as a valuable asset that has facilitated the rise and progression of novel fields, including data science, artificial intelligence (AI), deep learning (DL), and the internet of things (IoT). Effectively managing and analyzing data provides a competitive advantage for modern businesses. The objective of this study is to forecast the stock price of Turkish Airlines (THY), a publicly traded corporation listed on Borsa Istanbul. In order to achieve the intended objective, the utilization of machine learning approaches like SVM and XGBoost, as well as the deep learning algorithm Long Short-Term Memory (LSTM), are used. The models are trained over a time period including daily data from January 4, 2010 to September 5, 2023. The forecast performance of the models is evaluated by comparing the actual and predicted stock prices and the model with the lowest error is identified. The proposed models' performances are assessed using the RMSE, MSE, MAE, and R2 error statistics. According to the results obtained, it is determined that the LSTM model has lower error coefficients than SVM and XGBoost models and gives the best performance.

Key words: LSTM, stock price prediction, machine learning, deep learning, SVM, XGBoost.

Makine Öğrenimi ve Derin Öğrenme Algoritmalarını Kullanarak Hisse Senedi Fiyat Tahmini: Havacılık Sektörüne Yönelik Bir Örnek Çalışma

Öz: Teknolojik ilerlemelerle birlikte, insanlar çeşitli elektronik cihazlar ve sensörler aracılığıyla sürekli olarak veri üretmekte ve bu veriler dijital ortamlarda depolanmaktadır. Bu büyük veri havuzu, yeni disiplinlerin doğmasına ve gelişmesine olanak tanıyan bir kaynak haline gelmiş; örneğin, veri bilimi, yapay zekâ, derin öğrenme ve nesnelerin interneti gibi alanlar ortaya çıkmıştır. Verilerin etkili bir şekilde yönetilmesi ve analiz edilmesi, modern işletmeler için rekabet avantajı sağlamaktadır. Bu çalışma, Borsa İstanbul'da (BIST) işlem gören Türk Hava Yolları AO (THYAO) şirketinin hisse senedi fiyatının tahmin edilmesini amaçlamaktadır. Bu amaçla, makine öğrenmesi algoritmalarından Support Vector Machine (SVM) ve Extreme Gradient Boosting (XGBoost) ile derin öğrenme algoritması olan Long Short-Term Memory (LSTM) kullanılmıştır. Modeller, 4 Ocak 2010 ile 5 Eylül 2023 tarihleri arasındaki günlük verileri içeren bir zaman diliminde eğitilmiştir. Gerçek hisse senedi fiyatları ile tahmin edilen fiyatlar karşılaştırılarak modellerin performansları değerlendirilmiş ve en düşük hataya sahip model belirlenmiştir. Önerilen modellerin performansları RMSE, MSE, MAE ve R2 hata istatistikleri kullanılarak değerlendirilmiştir. Elde edilen sonuçlara göre LSTM modelinin SVM ve XGBoost modellerine göre daha düşük hata katsayılarına sahip olduğu ve en iyi performansı verdiği belirlenmiştir.

Anahtar Kelimeler: LSTM, hisse denedi fiyat tahmini, makine öğrenmesi, derin öğrenme, SVM, XGBoost.

1. Introduction

The constant scientifical and technological advancements has led to the perpetual generation of data by individuals through a multitude of electronic gadgets and sensors. Subsequently, this data is subsequently kept within digital surroundings. The substantial reservoir of data has served as a valuable asset, facilitating the rise and progression of novel fields of study, including including data science, artificial intelligence (AI), deep learning (DL), and the internet of things(IoT). In that point of view, data plays a pivotal part in various aspects of the corporate realm, encompassing the comprehension of customer behavior, product development, security protocols, tailored services, predictive analysis, and numerous other uses. Hence, proficiently overseeing and evaluating data confers a competitive edge to contemporary enterprises. The nature of data might vary in terms

* Sorumlu yazar: yegur@firat.edu.tr. Yazarın ORCID Numarası: [1] 0000-0001-6530-0598.

of its types and structures, contingent upon the specific source from which it is acquired. The data is organized into graphs and presented in a sequential manner. Sequential data refers to a collection of actions that are executed either by people or machines. Sequential data refers to data that is arranged in a specific manner based on a particular attribute or characteristic, and is presented in a systematic order. An illustration of this concept can be seen in the distinction between the stock market index, which represents time-ordered data, and genomic data, which is structured based on a specific rule. Temporal data, which is characterized by its time-dependent nature, possesses a temporal component and should be regarded as a chronological sequence. Data of this nature are commonly referred to as time series or sequential data [1].

Yet, the stock market is a financial environment that frequently exhibits intricacy, unpredictability, and non-linearity. The analysis of macroeconomic and microeconomic issues holds significant importance in investment decision-making within the dynamic structure of financial markets. Nevertheless, due to the dynamic nature of these components and their susceptibility to several unclear variables, it becomes very challenging to consolidate all macro and micro aspects and accurately ascertain the exact magnitude of their impacts. Hence, financial time series forecasting is widely acknowledged as an intricate domain of study within the realm of finance and investing. Besides macroeconomic indicators, certain variables at the micro-level within sectors can also impact stock values. During periods of economic uncertainty, precise forecasting of these values becomes crucial for investors in order to mitigate financial risk [2,3].

Turkish Airlines (THY), the national carrier of Turkey, operates from its headquarters in Istanbul and garners significant attention from international investors. Its inclusion among the top five corporations with the highest trading volume on Borsa Istanbul (BIST) demonstrates interest in this sector. The business structure of THY has a notable impact on the aviation sector, primarily in relation to oil prices and foreign exchange rates. The impact of oil as the primary cost component for airline firms, along with the procurement of essential equipment like airplanes in foreign currency, has significant implications for THY's cost structure and overall profitability. Hence, fluctuations in oil prices and foreign exchange rates possess the potential to exert a direct impact on the stock performance of THY. Simultaneously, the inclusion of THY in both the BIST 100 and the Transportation Index implies that fluctuations in these indices possess the potential to impact the stock prices of THY. Hence, it is vital to take into account these aspects while attempting to assess the stock valuation of THY, the foremost publicly traded corporation in Turkey.

The primary objective of this study is to predict stock price of Turkish Airlines (THYAO) within the context of its trading on Borsa Istanbul. To get the desired goal, the utilization of machine learning techniques such as XGBoost, along with the deep learning technique known as LSTM, are implemented. The models underwent training utilizing daily data including the time period from January 4, 2010 to September 5, 2023. The evaluation of model performance involved an examination of the actual stock values and their associated predicted prices. The model that exhibited the least amount of error was identified as the most optimal choice.

The forthcoming sections of this paper will offer a thorough evaluation of the current body of work on machine learning (ML) and deep learning (DL) techniques employed within the context of stock price prediction. The subsequent sections offer a thorough elucidation of the SVM, XGBoost, and LSTM approachs. The analysis phase of the study focuses on the examination of the implementation process, while the conclusion section evaluates the gathered results and offers recommendations.

2. Literature Review

This section presents a summary of studies that have employed SVM, XGBoost, and LSTM algorithms in the context of predicting stock prices.

In their study, Fenghua et al. utilized the SSA approach to analyze and evaluate stock prices. The utilization of the SSA methodology is implemented to split stock values into multiple components, encompassing trend, market fluctuations, and noise, over diverse time intervals. This methodology successfully identifies and separates the distinct components that represent these varied economic characteristics. The aforementioned characteristics are subsequently included with the SVM machine learning technique to produce predictions for stock prices. The study involves conducting tests to compare two unique strategies, namely "EEMD-SVM" and "SSA-SVM". These techniques aim to improve stock price forecasts by integrating price information into SVM algorithms. Based on existing research, it is indicated that the incorporation of pricing characteristics into SVMs has the capacity to improve the precision of predictions. The SSA-SVM approach has been found to produce the most precise prediction results [4].

In their study, Pawar et al. examined the utilization of artificial neural network models, namely RNN and LSTM, for the purpose of anticipating stock market trends and managing investment portfolios. The research is

carried out utilizing historical stock data of the stocks contained within the portfolio, with a specific emphasis on the analysis of time series. Moreover, this study conducts a comparative examination of RNN and LSTM models, in conjunction with traditional Machine Learning Algorithms such as Regression, SVM, Random Forest, Feed Forward Neural Network, and Back Propagation. The results of the paper indicated that the RNN-LSTM model has a greater level of accuracy when compared to traditional machine learning methods [5].

In their study, Yang et al. undertook research with the objective of providing accurate predictions of stock prices, thereby assisting investors in capitalizing on transient fluctuations in the market. The dataset utilized in this investigation was given by Jane Street. The dataset under consideration encompasses a substantial volume of data, encompassing anomalous instances such as missing data. Hence, the initial step involves conducting feature engineering on the dataset and applying averaging techniques to handle missing data. This procedure results in preprocessed data that is suitable for subsequent modeling purposes. The empirical findings indicate that the amalgamation of XGBoost and LightGBM in a hybrid model yields superior predictive capabilities compared to both individual models and the neural network. This paper highlights the significance of employing models such as XGBoost and LightGBM for the purpose of forecasting stock prices. Furthermore, it asserts that superior prediction outcomes may be achieved through the integration of these models [6].

The primary emphasis of the work conducted by Kanakam et al. revolves around the application of MLP and SVM machine learning techniques for the purpose of predicting stock market trends. The focus of this research is to employ an analysis of a certain company's stock prices in order to forecast future stock values. The predictive capabilities of the model are derived on an analysis of historical data, specifically the past stock prices. The study involved the analysis of around 1300 stock prices pertaining to a specific company. Multiple linear regression and SVM machine learning techniques were employed to forecast the present stock prices of the company using this dataset. According to the regression findings of the study, the accuracy of predicting the stock price using multiple linear regression was 99% and the accuracy of the model using SVM was 93% [7].

Tokmak utilized deep learning methodologies, notably Long-Short Term Memory networks, to forecast stock values in their research study. The focus of the study was around four specific stocks that are part of the Technology Index of Borsa Istanbul. A dataset consisting of a total of 2578 daily data points collected between the years 2012 and 2022 was constructed for the purpose of this study. The dataset was employed to carry out training and testing procedures using the model. According to the results, the testing procedure demonstrate that the forecasts demonstrate a significant level of consistency and a strong alignment with the actual events. The focus of this study revolves around the forecasting of stock prices for four specific stocks within the Technology Index, utilizing deep learning methodologies like as LSTM . The results of this investigation illustrate the reliability of these predictions [3].

Vuong et al. employed sophisticated machine learning and deep learning techniques to enhance the efficacy of the Stock Price Prediction system for both stock and Forex datasets. In this study, XGBoost was employed as a feature selection strategy to extract significant features and eliminate redundant features from time series data with a high dimensionality. The chosen characteristics are inputted into a deep LSTM network in order to predict stock values. The deep LSTM network employed to capture the spatial patterns inherent in the input time series and effectively leverage future contextual information. The experimental findings obtained by analyzing Forex data indicate that this particular strategy exhibits superior performance compared to the basic autoregressive integrated moving average approach, as evidenced by lower values of MAE, RMSE and MSE [8].

In their research, Kaneko and Asahi employed SVMs as a predictive model to forecast the future movement of the Nikkei index over three different time horizons: one day, one week, and one month. The study employed the historical rates of change of US stock prices and the Nikkei Stock Average as explanatory variables. Based on the findings of the investigation, it was determined that the predictive precision of the mean price of the Nikkei index for the subsequent day exhibits persistent enhancement in comparison to random forecasting [9].

The present investigation by Gülmez examined the application of the Artificial Rabbits Optimization (ARO) method in combination with LSTM, a deep learning model. Objective of this reseach was to examine the utilization of this integrated methodology in the field of stock price forecasting. The research employed the stock price data of the Dow Jones Industrial Average (DJIA) index in order to carry out the forecasting procedure. The study performed a comparative analysis of the LSTM-ARO model in comparison to other models to evaluate its performance. The supplementary models consist of an artificial neural network (ANN) model, three separate LSTM models, and an LSTM model enhanced through the utilization of a Genetic Algorithm (GA). The performance of the models was compared by a comparative study, which involved the examination of several metrics such as MSE, MAE, MAPE and R2. The study's results revealed that the LSTM-ARO model had a higher level of predictive ability in comparison to the alternative models [10].

İlkçar utilized machine learning methodologies to predict the stock prices of Turkish Airlines in their research. The research utilized a range of deep learning methodologies, such as LSTM, Feedforward Neural Network (FNN), and Gated Recurrent Unit (GRU). The study encompassed the training and evaluation of diverse neural network architectures specifically engineered to capture and represent both short-term and long-term memory functionalities. The evaluation of the models' performance was conducted by utilizing various metrics, including R-squared, MSE, RMSE and MAE. The results suggest that the system demonstrates a performance of 97% for FNN, and 99% for LSTM and GRU, as evaluated using the test R-squared performance metrics. The results emphasize the potential of machine learning as a valuable tool for enhancing decision-making processes in the prediction of sequential data sets. Moreover, the research results suggest that machine learning methods, including FNN, LSTM, and similar algorithms, demonstrate considerable potential in accurately forecasting indices related to the air transportation sector [1].

A study by Almaafi et al. compared and evaluated the ARIMA and XGBoost models' forecasting capabilities for Saudi Telecom Company's weekly closing stock prices. The study's findings indicate that the XGBoost model outperformed the ARIMA model in terms of all evaluation measures. The findings demonstrate how well machine learning techniques predict stock market prices. Furthermore, this serves as an illustration of the limitations present in traditional statistical models when trying to predict fluctuations in stock prices, while highlighting the capability of machine learning techniques to reveal hidden links and patterns in datasets [11].

Dezhkam and Manzuri developed a novel model called "HHT-XGB" for predicting the closing values of stocks in the future. The suggested model incorporates the Hilbert-Huang Transform (HHT) for feature engineering and utilizes XGBoost as a classifier to identify close price trends. The categorization output represents a rating system that assesses the performance of stocks, enabling the optimization of portfolio weights for equities exhibiting superior trading performance. The study's findings indicate that the portfolios optimized within the scope of this research had superior performance compared to the portfolios designed only based on raw financial data, with a notable outperformance of 99.8%. In addition, empirical analysis indicates that the HHT-XGB strategy exhibits superior performance compared to benchmark strategies, especially in periods of market underperformance [12].

3. Data and Methodology

In this study, 3433 days of data between 01.01.2010-05.09.2023 of Turkish Airlines Corporation (THYAO), the largest aviation company in Turkey, were used. The data were obtained from www.investing.com. The time series graph of stock price data is given in Figure 1.

Figure 1. Time series of THYAO.

This study assesses the predictive performance of the proposed LSTM, XGBoost, and SVM models by evaluating metrics such as MSE, RMSE, MAE and R2. Equations 1, 2, 3, and 4 provide the appropriate mathematical expressions for the MSE coefficient, RMSE coefficient, MAE coefficient, and R2 coefficient, which are among the coefficients considered in this research.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{N}} \quad (2)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \quad (3)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2} \quad (4)$$

3.1. LSTM model

Recurrent neural network-based LSTM models are well known for their efficiency at learning data with time-series characteristics or sequences [13,14]. An artificial neural network called an LSTM network uses LSTM units as its main building blocks, either in place of or in addition to other network units. The LSTM unit is a type of recurrent network unit that exhibits exceptional proficiency in retaining information over extended or brief time intervals. The crucial aspect of this capability is in the utilization of recurrent components that do not incorporate any activation function. Therefore, it can be observed that the stored value does not undergo iterative compression over a period of time, resulting in the gradient or blame term not tending to diminish when Backpropagation via time is employed for training purposes [15].

The normalization of the data sets was performed using the min-max method, as outlined in Equation (5), in order to establish a standardized range of values between 0 and 1.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (5)$$

The forget gate in the LSTM approach is in charge of specifically discarding the previous sequence's cell state. The activation function σ_g processes the time series' current input, denoted by the symbol x_t, as well as its previous hidden state, denoted by the symbol h_{t-1}. The output vector f_t, which is related to the forget gate, is produced as a result of this processing. Equation (6) can be used to express this relationship mathematically.

$$f_t = \sigma_g\,(W_f x_t + U_f h_{t-1} + b_f) \quad (6)$$

The bias coefficient is commonly denoted as b_f, whereas the weight coefficients associated with the forget gate are typically referred to as W_f and U_f. The activation function is denoted by the sign σ_g. The responsibility of the input gate is to regulate the influx of input data that is directed towards the active cell during a specific period. The coefficientsi_t and C'_t in this gate are determined by the current time series input x_t and the hidden state h_{t-1} from the previous time step. The utilization of coefficients is crucial for the preservation of cell candidate data and the calculation of the proportion of information requiring updates. The utilization of the activation function is employed in the computation of these coefficients. The symbols W_i, U_i, W_c, and U_c represent the weight coefficients. The bias coefficients in Equations (7) and (8) are denoted as b_i and b_c., respectively. The activation functions are symbolized by σ_g and σ_c.

$$i_t = \sigma_g\,(W_i x_t + U_i h_{t-1} + b_i) \quad (7)$$

$$C'_t = \sigma_c\,(W_c x_t + U_c h_{t-1} + b_c) \quad (8)$$

The process of updating the cell state in a linear interaction approach entails two steps: first, updating the prior cell state, and second, combining this updated prior cell state into the present cell state to complete the information update. To be more precise, the cell state, represented as C_t, undergoes an update process where it is obtained by multiplying the output of the forget gate, f_t, with the previous cell state, C_{t-1}, and adding the product

of the output of the input gate, i_t, and the cell candidate data, C'_t. The computation provides the description of the updated cellular state, C_t, as stated in Equation (9).

$$C_t = f_t \, x \, C_{t-1} + i_t \, x \, C'_t \tag{9}$$

The computation and data transfer to the following time step are under the control of the output gate. Equation (10) illustrates how the activation function o_g is applied to the input vectors h_{t-1} and x_t to produce the output vector o_t.The output gate controls the computation and data transfer to the subsequent time step. Equation (10) demonstrates the application of the activation function o_g to the input vectors h_{t-1} and x_t, resulting in the generation of the output vector o_t.

$$o_t = \sigma_g \, (W_o x_t + U_o h_{t-1} + b_o) \tag{10}$$

The input gate is linked to the weight coefficients of the cell state, W_o and U_o, along with the bias coefficient, c_o. After that, the output gate produces the output o_t, which is subsequently multiplied by the current sequence cell state c_t. The outcome is subsequently subjected to the activation function tanh in order to produce the ultimate output of the concealed layer, as indicated by Equation (11).

$$h_t = o_t \text{ x } tan\, h\, (C_t) \tag{11}$$

3.2. XGBoost model

Chen and Guestrin developed the XGBoost method in 2016 [16]. The proposed method is derived from the Classification and Regression Tree (CART) algorithm. It introduces a novel approach to redefine the partition attributes and employs the minimization of the loss function to decide these attributes. One notable benefit of employing these methodologies is that the forecast is derived from empirical data. Boosting techniques have been found to yield efficient outcomes when used to regression and classification trees. XGBoost has demonstrated efficacy in generating accurate predictions and offers computational advantages over many alternative machine learning approaches. According to Abar, this feature possesses the capability to be utilized efficiently in various applications such as classification, consumer behavior prediction, motion detection, and advertising multidimensional big data analysis [17]. In order to predict the outcome for a dataset with n samples and m features, a tree ensemble model uses K additive functions $D = \{(x_i, y_i)\}(|D| = n, x_i \in \mathcal{R}^m, y_i \in \mathcal{R}$.

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k(y_i), f_k \in \mathcal{F} \tag{12}$$

In Equation (12), $\mathcal{F}$ represents the space of regression trees , f_k is the quantity of weak learners, and K signifies the total count of weak learners.

$$\min L^{(t)}(y_i, \hat{y}_i^{(t)}) = \min\left(\sum_{i=1}^{n} \iota\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k)\right) \tag{13}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda w^2 \tag{14}$$

The algorithm's objective function at time t, denoted as $L^{(t)}$, is defined by Equation (13). The parameter $l\left(y_i, \hat{y}_i^{(t)}\right)$ represents many forms of loss functions that are employed to address certain challenges. It is commonly utilized to quantify the extent of discrepancy between the actual (y_i) and the predicted ($\hat{y}_i^{(t)}$) value and $\sum_{k=1}^{t} \Omega(f_k)$ as a gauge of the overall complexity of the model, as presented in Equation (14).

$$\min L^{(t)} = min\left(\sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)\right] + \Omega(f_t)\right) \tag{15}$$

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, y_i^{t-1}) \tag{16}$$

$$h_i = \partial^2_{y_i^{t-1}} l(y_i, y_i^{t-1}) \tag{17}$$

The objective function is evaluated by substituting the predicted ($\hat{y}_i^{(t)}$) of the ith sample in the tth iteration. The calculation is performed using the second-order approximation of the Taylor expansion at $\hat{y}_i^{(t-1)}$, as shown in Equation (15). In this, the variables g_i and h_i represent the first and second derivatives of the loss function $l(y_i, \hat{y}_i^{(t)})$, correspondingly.

$$w_j^* = -\frac{\sum g_i}{\sum h_i + \lambda} \tag{18}$$

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum g_i)^2}{\sum h_i + \lambda} + \gamma . T \tag{19}$$

By substituting Equation (15), Equation (16), and Equation (17) into Equation (13), we may proceed to take the derivative. Solutions can be derived from Equations (18) and (19). Equations (18) and (19) denote the variable obj^*, which corresponds to the value of the loss function's score. A lower score indicates a more optimal tree structure. The symbol w_j^* denotes the solution of weights in the context being discussed.

3.3. SVM model

SVM algorithm is founded around the estimation of an optimal discriminative function for data classification, utilizing either a linear or non-linear function. The primary goal of SVM is to identify the ideal decision border that optimizes the margin between the boundaries of the different data classes. The method achieves this objective by detecting data points known as support vectors, which establish the boundaries for each class. SVM exhibit resilience to outliers and provide strong performance in feature spaces with a high number of dimensions. The kernel functions that are frequently employed in many applications encompass linear, nonlinear, polynomial, Gaussian kernel, radial basis function (RBF), and sigmoid [18].

$$min\frac{1}{2}\left|\left|w^2\right|\right| + C\sum_{i=1}^{n}(\zeta_1 + \zeta_i^*) \tag{20}$$

$$(w\emptyset(x_i) + b) - y_1 \leq \varepsilon + \zeta_1 \tag{21}$$

$$y_t - (w\emptyset(x_i) + b) \leq \varepsilon + \zeta_i^* \tag{22}$$

In this context, the symbol w is used to indicate a direction vector, while "C" represents an adjustment factor. The variables , ζ_i and ζ_i^*∗ are referred to as slack variables. The function $\phi(x_i)$ is used to transfer the input vector x_i to a high-dimensional hyperspace. The symbol "b" represents the intercept of the regression function, while ε is a coefficient that denotes non-sensitivity. The initial component of the objective function represents the level of complexity within the model, whereas the subsequent component represents the degree of error in fitting. In the theory of SVMs, the model achieves optimal performance when the sum is minimized. SVM models aim to achieve the optimal balance between the generalization performance of the model and its fitting performance.

The SVM model can be formulated as a quadratic programming problem in large dimensions. In order to mitigate the occurrence of a "dimensional disaster," the implementation of a kernel function is proposed as a means to transform high dimensional computational processes into low dimensional ones. Various types of kernel functions, including linear, RBF, Gaussian, polynomial, and other kernel functions, are frequently utilized. When processing high-dimensional complex samples, the RBF kernel outperforms the linear kernel. The parameterization of the RBF kernel function is also much simpler than that of the Gaussian and polynomial kernel functions. When choosing a solution for the SVM model, the RBF kernel is usually chosen as follows:

$$k(x, x_i) = e^{-g\|x - x_i\|^2} \tag{23}$$

The parameter g is employed to calibrate various distributions of samples.

4. Results and Discussion

The findings of the suggested LSTM, XGBoost, and SVM models are presented in Table 1. The present research utilized four statistical metrics. The evaluation tool employed to quantify the discrepancy between projected values and actual values is commonly known as RMSE. A lower score indicates a higher level of alignment between the model's predictions and actual observations. The mean of the squared deviations between

the expected and actual values is computed for calculating the MSE. The model's predictions exhibit more accuracy as the MSE decreases. The average magnitude of the variances between expected and actual values is quantified by the MAE metric. When the model's predictions are highly accurate, as indicated by a low mean absolute error (MAE), the impact of extreme data points is minimized. The fraction of the dependent variable's variance that can be quantified by the model is measured by the R2 statistic. When the coefficient of determination (R2 value) becomes close to 1, it means that the model can explain a sizable percentage of the observed variability.

Table 1. Statistical results of the models.

	SVM	**XGBOOST**	**LSTM**
RMSE	1.5925	1.5887	0.02431
MSE	2.5361	2.5241	0.00059
MAE	0.4796	0.4959	0.01572
R2	0.9983601	0.998368	0.990854

The outcomes of the performance criteria used to assess three different machine learning models are presented in Table 1. After doing a thorough examination of the data, it is evident that the SVM model demonstrates the subsequent numerical outcomes: RMSE of 1.5925, MSE of 2.5361, MAE of 0.4796, and an R-squared (R2) value of 0.9983601. In this particular instance, the root mean square error (RMSE) of 1.5925 signifies that the SVM model's predictions exhibit an average deviation of around 1.59 units from the actual values. The aforementioned figure signifies a significantly diminished magnitude of inaccuracy. The MSE score of 2.5361 signifies that, on average, the model's predictions exhibit a squared departure of 2.54 units from the actual data. MAE quantifies the average variation between predictions and true values, employing the absolute values of these variances. MAE score of 0.4796 suggests that the SVM model's forecast results exhibit an average variation of around 0.48 units from the actual values. This indicates a significant degree of mistake at a minimal level. The coefficient of determination (R2) quantifies the extent to which the model elucidates the variability in the independent variables. The R-squared value of 0.9983601 suggests that the SVM model has a high degree of accuracy in capturing the patterns present in the data, and effectively accounts for nearly all of the variability observed in the independent variables. The obtained R2 value is strong, suggesting that the model effectively accounts for the observed data.

The root MSE, RMSE of XGBoost, which is 1.5887, suggests that, on average, the model's forecasts exhibit a deviation of around 1.59 units from the true values. This indicates a relatively minimal degree of error. The MSE score of 2.5241 signifies that the model's predictions exhibit an average squared deviation of 2.54 units from the actual data. The R-squared value of 0.998368 suggests that the XGBoost model has a high degree of accuracy in fitting the data, effectively accounting for nearly all of the variability observed in the independent variables. The high R2 value observed in this analysis suggests that the model effectively accounts for the variability in the data.

RMSE value of the LSTM model, which is 0.02431, signifies that the average deviation between the predicted and the actual values is merely 0.024 units. The MSE value of 0.00059 suggests that, on average, the model's forecasts exhibit a deviation of approximately 0.00059 units squared from the actual values. MAE score of 0.01572 signifies that the LSTM model's predictions exhibit an average variance of merely 0.015 units from the actual data. The R2 value of 0.990854 suggests that the LSTM model exhibits a high degree of accuracy in capturing the patterns present in the data and effectively accounts for a significant percentage of the variation in the independent variables. The strong R2 value suggests that the model effectively accounts for the variability observed in the data.

Upon doing a comprehensive analysis of Table 1, it is evident that the SVM model demonstrates a notable proficiency in minimizing errors across many metrics, including RMSE, MSE, MAE. The coefficient of determination R2 exhibits a remarkably high value, indicating a strong explanatory power in relation to the observed data. This showed that the SVM model has a high level of performance. The performance metrics of the XGBoost model is comparable to that of the SVM. The root MSE, RMSE, MSE, and mean absolute error (MAE) exhibit somewhat elevated values, while the R-squared R2 number demonstrates a substantial degree of explanatory power, indicating a strong fit to the data. The LSTM model demonstrates a notable reduction in errors when evaluated using the root MSE, RMSE, MSE, and MAE metrics. Nevertheless, the R2 value of the

aforementioned model is marginally inferior to that of the SVM and XGBoost models, indicating a slightly lesser degree of explanatory power in relation to the data. In order to determine the most effective model, it is necessary to take into account all of the metrics. The coefficient of determination (R2) provides a measure of the extent to which the model explains the observed data. On the other hand, RMSE, MSE and MAE serve as indicators of the accuracy of the model's predictions. In this analysis, all three models demonstrate relatively high R2 values, indicating a strong level of explanatory power in relation to the observed data. Nevertheless, the LSTM model demonstrates superior predictive accuracy compared to other models, as seen by significantly fewer errors in metrics such asRMSE, MSE, and MAE. Consequently, the LSTM model demonstrates superior predictive performance as evidenced by its notably lower values of RMSE, MSE, and MAE, as derived from the available data.

Based on the findings of the examination, it is seen that the LSTM model demonstrates superior performance when evaluating the performance measures, despite a slightly lower R2 value. Figure 2 depicts the extent to which the proposed LSTM model's predictions align with the observed data. This study applies LSTM, XGBoost, and SVM models to analyze the THYAO stock traded in BIST100. The study utilized a total of 3433 daily data points of THYAO between the dates of April 1, 2010, and May 9, 2023. The assessment of the prediction capabilities of the suggested models involved the utilization of statistical measures such as RMSE, MSE, MAE, and R2. The LSTM model demonstrates superior performance across all error statistics when evaluating the derived performance measures.

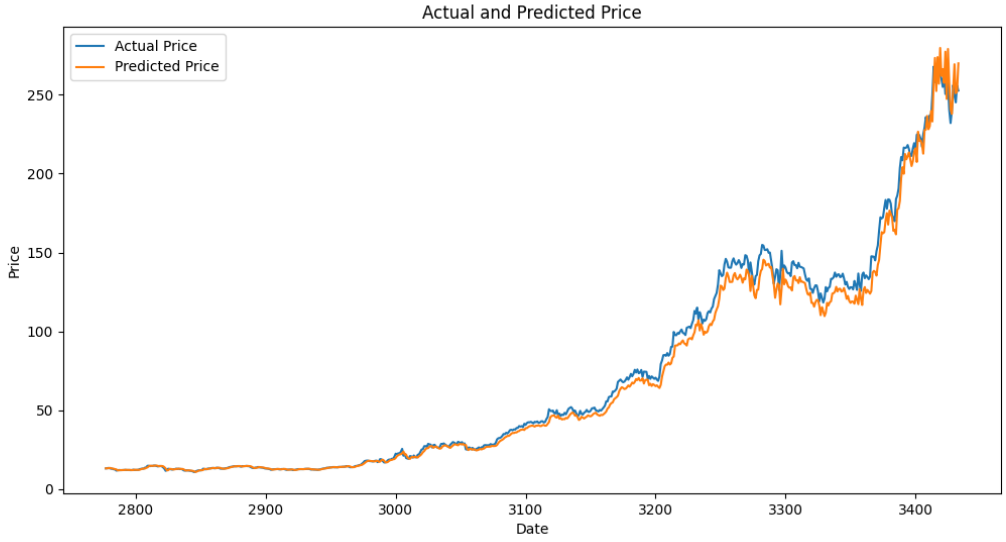

Figure 2. Actual and predicted values for LSTM model

As a result, it can be said that when it comes to accurately forecasting the values of publicly traded companies on stock exchanges, the LSTM model outperforms the SVM and XGBoost models. The study's conclusions will produce useful recommendations that will be advantageous to both individual and institutional investors. In prospective researchs, it is possible to examine the outcomes derived from diverse time series, machine learning, and deep learning algorithms, including ARIMA, CNN, RBF, MLP, RF, and Decision Trees. The suggested approach has the potential for generalization through its application to various data sets. Furthermore, it should be noted that this study exclusively utilizes the THYAO stock price as the sole variable in the model. In prospective investigations, supplementary factors such as interest rates, inflation, and currency rates, which exert an influence on stock prices, may be incorporated.

References

[1] İlkçar, M. (2023). Turkish Airlines BIST share price prediction with deep artificial neural network considering trading volume and seasonal values. International Journal of InformaticsTechnologies, 16(1), 43-53.

[2] Çınaroğlu, E, Avcı, T. (2020). Prediction of THY stock value with artificial neural networks. Atatürk University Journal of Economics and Administrative Sciences, 34(1), 1-19.

[3] Tokmak, M. (2022). Stock price prediction using Long-Short-term memory network. Mehmet Akif Ersoy University Journal of Applied Sciences, 6(2), 309-322.

[4] Fenghua, WEN, Jihong, XIAO, Zhifang, HE, Xu, GONG. (2014). Stock price prediction based on SSA and SVM. Procedia Computer Science, 31, 625-631.

[5] Pawar, K, Jalem, RS, Tiwari, V. (2019). Stock market price prediction using LSTM RNN. In Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018 (pp. 493-503). Springer Singapore.

[6] Yang, Y, Wu, Y, Wang, P, Jiali, X. (2021). Stock price prediction based on xgboost and lightgbm. In E3s web of conferences (Vol. 275, p. 01040). EDP Sciences.

[7] Kanakam, R, Ramesh, D, Mohmmad, S, Shabana, S, Prakash, TC. (2022, May). Stock price prediction using multiple linear regression and support vector machine (regression). In AIP Conference Proceedings (Vol. 2418, No. 1). AIP Publishing.

[8] Vuong, PH, Dat, TT, Mai, TK, Uyen, PH. (2022). Stock-price forecasting based on XGBoost and LSTM. Computer Systems Science & Engineering, 40(1).

[9] Kaneko, T, Asahi, Y. (2023). The Nikkei Stock Average Prediction by SVM. In International Conference on Human-Computer Interaction, 211-221.

[10] Gülmez, B. (2023). Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm. Expert Systems with Applications, 227, 120346.

[11] Almaafi, A, Bajaba, S, Alnori, F. (2023). Stock price prediction using ARIMA versus XGBoost models: the case of the largest telecommunication company in the Middle East. International Journal of Information Technology, 15(4), 1813-1818.

[12] Dezhkam, A, Manzuri, MT. (2023). Forecasting stock market for an efficient portfolio by combining XGBoost and Hilbert–Huang transform. Engineering Applications of Artificial Intelligence, 118, 105626.

[13] Schuster, M, Paliwal, K. (1997), Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 1997, 45, 2673–2681.

[14] Hochreiter, S, Schmidhuber, J. (1997), Long Short-Term Memory. Neural Comput. 1997, 9, 1735–1780.

[15] Chen, X, Wei, L, Xu, J. (2017). House Price Prediction Using LSTM. http://arxiv.org/abs/1709.08432

[16] Chen, T, Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu, 785–794. https://doi.org/10.1145/2939672.2939785

[17] Abar, H. (2020). Estimation of Gold Prices by Xgboost and Mars Methods. Ekev Academy Journal, (83), 427-446.

[18] Bakiler, H. (2023). Classification of gases with deep network based attributes and regression analysis of concentration values. Başkent University Institute of Science and Technology Unpublished Doctoral Thesis,2023