SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
VOL. 5, NO. 3, DECEMBER 2022
DOI: 10.35377/saucis.05.03.1200151

Research Article

Sector-Based Stock Price Prediction with Machine Learning Models

Doğangün Kocaoğlu[1], Korhan Turgut[2], Mehmet Zeki Konyar[3]
[1]Kocaeli University, Department of Computer Engineering; dogangun82@gmail.com
[2]Kocaeli University, Department of Computer Engineering; ateshan@yahoo.com
[3]Corresponding Author; Kocaeli University, Department of Software Engineering; mzeki.konyar@kocaeli.edu.tr

Received 7 November 2022; Accepted 30 November 2022; Published online 31 December 2022

Abstract

Stock price prediction is an important topic for investors and companies. The increasing effect of machine learning methods in every field also applies to stock forecasting. In this study, it is aimed to predict the future prices of the stocks of companies in different sectors traded on the Borsa Istanbul (BIST) 30 Index. For the study, the data of two companies selected as examples from each of the holding, white goods, petrochemical, iron and steel, transportation and communication sectors were analyzed. In the study, in addition to the share analysis of the sectors, the price prediction performances of the machine learning algorithm on a sectoral basis were examined. For these tests, XGBoost, Support Vector Machines (SVM), K-nearest neighbors (KNN) and Random Forest (RF) algorithms were used. The obtained results were analyzed with mean absolute error (MAE), mean absolute percent error (MAPE), mean squared error (MSE), and R^2 correlation metrics. The best estimations on a sectoral basis were made for companies in the Iron and Steel and Petroleum field. One of the most important innovations in the study is the examination of the effect of current macro changes on the forecasting model. As an example, the effect of the changes in the Central Bank Governors, which took place three times in the 5-year period, on the forecast was investigated. The results showed that the unpredictable effects on the policies after the change of Governors also negatively affected the forecast performance.

Keywords: Borsa Istanbul, machine learning, stock price prediction,

1. Introduction

Persons and institutions investing in the capital market should know and follow the market in which they invest. Therefore, all individual and institutional investors are required to make market forecasts by providing accurate and fast all economic and financial information about the general economy, sectors and the institutions they invest in. However, the difficulty of predicting people's feelings and expectations reduces the chances of any analysis system that can be considered fully successful. In addition, the fact that the people who set the prices (market professionals, institutional investors, speculators, manipulators) have different cultural, educational and knowledge structures make the situation even more difficult. There are different methods for stock price analysis in the literature. Fundamental analysis, technical analysis and statistical forecasting methods are the most frequently used methods [1].

In fundamental analysis, which is the most comprehensive method used in the evaluation of stocks, the actual value of the stock is tried to be calculated by considering all possible factors that may affect the value of stocks. The factors that can affect the value of the stock can be grouped under three main titles: economic analysis, sector analysis and firm analysis. As a result of these analyzes, the risk and return relationship of the stock is revealed and its real value is calculated with the help of various methods. If the calculated real value is higher than the market value, the share is purchased; otherwise, it will not be traded. In fundamental analysis, there are stages such as economic analysis, sector analysis, firm analysis, risk and return estimation, respectively. The first stage of fundamental analysis is economic analysis. With this analysis, it is checked whether the general economic conjuncture is suitable for investing in stocks affected by general economic conditions. When a positive result is obtained from the economic analysis for stock investment, the second stage of the fundamental analysis, the industry analysis, is started. At this stage, it is tried to determine which of the many sectors in the economy should be invested in.

In technical analysis, it is aimed to determine the direction of the market and stock prices by using certain market data. Market data used in technical analysis consists of stock market index or price transaction volume information for stocks. Technical analysts try to predict the future direction of stock prices based on past market data. With statistical forecasting methods, time series analysis is used to predict the future, and it is tried to determine the attitudes of the series towards the future outside the forecast period. The traditional time series analyzes past data and tries to calculate its future approximation in the form of linear combinations of this historical data. In other words, a model is tried to be established in relation to the past values of the nonlinear values of a variable [2].

Professional knowledge and skills are very important for analyzes made with traditional methods. It is necessary to evaluate many parameters together and to read the behavior of the market from past to present. So, in recent years, it has become very popular to use artificial intelligence methods to make stock analysis and forecasting processes faster and easier. Models trained with the features in the datasets make successful predictions in the face of a situation they have never seen before. Modeling of stocks has become easier thanks to machine learning and deep learning algorithms [3],[4].

In this study, it is aimed to predict stock prices by machine learning method. In the literature, it has been seen that machine learning and deep learning algorithms such as Artificial Neural Networks, Random Forest, XGBoost, SVM, KNN and Long Short-Term Memory are used for prediction [5]. In this study, Random Forest, XGBoost, SVM and KNN, which are the most used machine learning algorithms, were used. It has been observed that the existing studies in the literature generally predict BIST-30 or BIST-100 index, and when stock-based prediction are made, stocks are randomly selected. The most important contribution of this study proposed in this article to the literature is to make a sector-based estimation. Another important contribution of our study is the analysis of the effect of some specific periods, such as the change of the Central Bank governors, on the stock price prediction with the machine learning.

The rest of the paper is organized as follows; section 2 the similar literature studies are summarized. Details of the machine learning methods are given in section 3. The proposed methods and experimental results are given in section 4. In the last section the conclusions are summarized.

2. The Related Studies

The increasing effect of machine learning methods in every field has also become important for stock forecasting. In this section, some of the existing machine learning and deep learning studies in the literature are summarized.

In the [6], the price estimates of the stocks of 5 Turkish Banks were made according to the stock market. By using various indicator values of the shares, estimation was made with decision tree, multiple regression, and random forest machine learning models. The success of the estimation results obtained was evaluated with the R^2 metric and values between 0.95-0.98 were reached. In the [7], the direction of change of the BIST 50 index was estimated with artificial neural networks (ANN), KNN, Naive Bayes and C4.5 decision tree models. The success of the estimation results obtained was evaluated with the classification accuracy, and the highest value was obtained as 92.71% with the C4.5 decision tree model. In the [8], the SP500 stock market index was estimated using a CNN-based forecasting model. In the study, an answer was sought on how to use the convolutional neural networks (CNN) model in stock market forecasting and how to optimize it. For the estimation process, 4 different CNN models based on different parameters were used. The obtained results were compared with the support vector machine (SVM) model and artificial neural network (ANN) model. In [9] ,daily returns of stocks in the Macedonian Stock Exchange were tried to be estimated based on linear regression and correlation analysis. The daily statistical forecast values of the stocks were evaluated over the R^2 metric.

In the study of [10], a forecasting system is proposed for stocks in NYSE, NASDAQ and NYSE MKT stock exchanges using deep learning models LSTM (Long Short-Term Memory), Gated Recurrent Unit (GRU) and Bidirectional LSTM. Experimental results were obtained with the BLSTM model with an accuracy of 63.54%. According to [11] a study was conducted on Coca-Cola Company shares. The study aimed to determine whether SVM is more accurate than Linear Regression. The estimation results were evaluated using Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute

Percent Error (MAPE), Mean Square Error (MSE) and Correlation Regression (R^2) and nonlinear multiple correlation factor evaluation criteria. As a result of the analysis, it was seen that SVM achieved more accurate results than LR. In the method of the [12], it is aimed to overcome the difficulty of forecasting crude oil prices due to the chaotic behavior of the time series. In the study, a wrapper-based feature selection approach using the multi-objective optimization technique and a support vector regression (SVR)-based prediction model are proposed. In the proposed model, features based on technical indicators such as simple moving average (SMA), exponential moving average (EMA) and Kaufman's adaptive moving average (KAMA) are used.

A data set was created by collecting the financial values of 22 companies traded on BIST-30 between the years 2010-2019 in the [13]. The stock prices of the companies were estimated using Artificial Neural Networks (ANN), Random Forest (RF) and XGBoost algorithms. The estimation results were compared over the R^2 metric, and the highest value was obtained with XGBoost as 0.758. In the study of [14], price prediction was made on the trading data on the Bitcoin stock market. The Linear Regression model was used for the estimation process and the results were evaluated over the R^2 metric. In the study of [15],stock price prediction was made by using machine learning and deep learning methods. Polynomial Regression, Random Forest Regression, Recurrent Neural Networks (RNN) and Long-Short-Term Memory (LSTM) methods were used for estimation. The estimation results were evaluated with RMSE, MAE, MSE metrics. The lowest error value was obtained with the Random Forest Regression model and the highest error value was obtained with the Polynomial Regression model. In addition to financial values, the authors of the [16] predicted stock market trends by making use of gold and oil prices. LSTM and CNN models were used to classify the data of SP500 index and compare investment returns. In experimental studies, the accuracy rate of the model increased up to 67%. In addition, it has been determined that this method will provide a return of around 13% with the investment. A 2-dimensional CNN-based forecasting approach is proposed for stocks in the Dow30 index proposed in the [17]. In the created rule-based model, the next day's buying, selling, or holding position of the stock is tried to be estimated.

Various machine learning algorithms were used for empirical asset pricing on the Chinese stock market in [18]. In the study, a comprehensive set of return estimation factors was created and analyzed. On the dataset, least absolute shrinkage and selection operator (LASSO), ordinary least squares (OLS) regression, partial least squares regression (PLS), gradient boosted regression trees (GBRT), elastic net (Enet), random forest (RF), variable subsampling Estimation aggregation (VASA), and neural networks methods based predicitons were made. The prediction performance of the models was examined through the R^2 metric in the predictions. In addition, predictability between different sub-samples was evaluated. In the [19], it is aimed both to predict the price of the stock and to compare the results obtained with Kalman filters, XGBoost and Auto Regressive Integrated Moving Average (ARIMA) models. We also compare the results of four models, including a hybrid model combining Kalman filters and XGBoost, to predict the price of New York Stock Exchange (NYSE) and National Stock Exchange (NSE) stocks. In the comparison, the lowest accuracy was obtained with the NYSE data set of the Kalman filter model as 64.96%. The highest accuracy was found to be 90.11% with the NYSE dataset of the XGBoost model.

3. The Machine Learning Methods

In this section, the machine learning methods used in the article will be briefly explained. In the article, four different regression methods were used for the stock analysis process. Fort he regression of the stock price prediciton Support Vector Machines (SVM), K-nearest neighbors (KNN), XGBoost and Random Forest (RF) based algorithms were used.

3.1 Support Vector Machine (SVM) Regression

Support Vector Regression is a regression model which uses the SVM approach that supports both linear and nonlinear regression operations. SVM based machine learning algorithm is used for both regression and classification problems. In the SVM, each data element is plotted as a point in the corresponding space such that the value of each attribute is a certain value in the coordinate system. Then, the classification is made by obtaining the hyperplane that best separates the two classes. Support Vector Regression works according to the SVM principle with minor differences [20]. The points given in the data are used to find the regression curve. Since the process is done with a regression algorithm, the curve obtained is not used as a decision limit, but to obtain the match between the vector and the curve. In normal regression, the aim is to minimize the error rate. In SVR, it is aimed to fit the error within a threshold. Therefore, the SVR model is used to estimate the best value for data in a given range.

3.2 K-Nearest Neighborhood (KNN) Regression

Although linear regression cannot provide very precise estimates of time, it is very useful for some critical problems. Thanks to the linear regression approach, many alternative models such as KNN have emerged that can be used in the field of machine learning. The KNN algorithm is a supervised learning algorithm in which a target variable is estimated using one or more independent variables [21]. Regression is the construction of a predictive function in which the target variable is numeric. Some algorithms can only classify, some can only regress, and some can do both classification and regression. The KNN algorithm adapts seamlessly to both classification and regression.

3.3 Extreme Gradient Boosting (XGBoost) Regression

XGBoost as a community learning method has been showing significant results recently. Relying on the results of a single machine learning model is not enough for some critical prediction and classification operations. Community learning offers more valuable results for combining the predictive power of multiple students. This result is achieved with a single model that consists of several models and gives a collective output. In the strengthening phase of the model, trees are created sequentially, so each subsequent tree aims to reduce the errors of the previous tree. Each tree learns new information from its predecessors, reducing existing errors. Therefore, a tree in a queue will learn from an updated version of what remains. The basic learners used for support are weak learners and their power for prediction is slightly better than random prediction. Each of weak learners contributes some critical information for prediction. Despite the weaknesses of these learners, by combining them effectively, the reinforcement technique reveals a powerful learning method. Parameters such as the number of trees or iterations, the depth of the tree can be optimally selected through validation techniques such as the learning rate of gradient boosting and k-fold cross validation [22].

3.4 Random Forest (RF) Regression

This regression is a collection of prediction trees based on the RF classifier. Each tree has a similar distribution to other trees in the random forest and depends on independently sampled random vectors. Random forest is an RF-based modeling technique used in behavior analysis and predictions. It contains several decision trees that represent a different classification example of random forest data entry. The random forest technique considers the samples one by one and takes the most voted sample as the selected prediction. Each tree in the classifications receives its inputs from the samples in the first dataset. Then, randomly selected features are used to grow the tree at each node. Each of the trees in the forest should not be pruned until the end of the exercise until the estimate is reached with certainty. In this way, the random forest ensures that any classifier with weak correlations will be a strong classifier. The random forest technique can also process big data with thousands of variables. A class can automatically balance datasets when data is less sparse than other classes[23].

4. The Proposed Method and Experimental Studies

The flow chart of the proposed method of this study for the analysis of stocks of different sectors given in Figure 1. First, the data set was collected and prepared by examining the data of two companies selected as examples from the Holding, White Goods, Petrochemical, Iron and Steel, Air Transport and Communication sectors. In Table 1, the company shares, and the sectors of the companies used for the experimental studies are given. The end-of-day closing prices of the stocks, which are the output of the models used in this study, are taken from the publicly available data on the website of IS Investment (https://www.isyatirim.com.tr) [24] for the last five-year period (01.10.2016 – 30.09.2021).

Table 1 Stock dataset information

Stock Name	Sector
SAHOL	Holding
KCHOL	Holding
EREGL	Iron and Steel
KRDMR	Iron and Steel
TUPRS	Petrochemical
PETKM	Petrochemical
ARCLK	White Goods
VESTL	White Goods
TRCELL	Communication
TTKOM	Communication
THYAO	Air Transport
PGSUS	Air Transport

The financial statement data used in the technical analysis of the companies were obtained from the website of IS Investment. Various firm ratios (current ratio, acid-test ratio, cash ratio, net profit margin) revealed by these data are used as inputs in our model. Macro data such as policy rate, inflation, and USD/TL parity used in the fundamental analysis of the companies were obtained retrospectively from institutional websites such as the Central Bank of the Republic of Turkey (TCMB) and the Turkish Statistical Institute (TUIK).

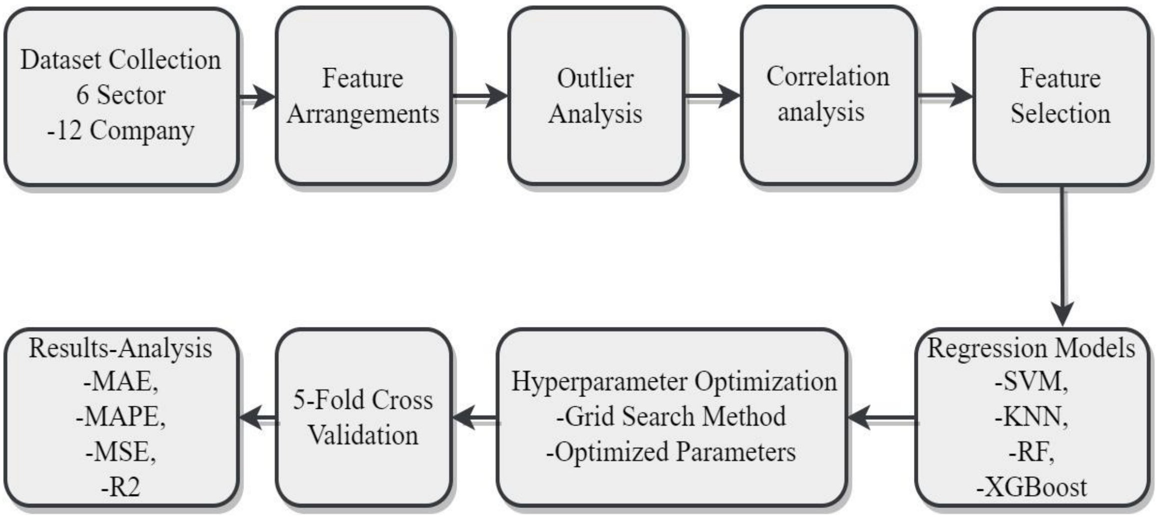

Figure 1 Flow chart of the proposed method

The frequency with which data is disclosed to supply the forecast model varies. Therefore, an imbalance arises in the training of machine learning models. To eliminate this imbalance, certain arrangements have been made to convert the data to the ones with the highest frequency. Since the interest rate data are shared weekly by the TCMB, the announced weekly rate has been retrospectively accepted as the same every day for seven days. Inflation rate data is shared monthly by TUIK, and the announced monthly rate has been accepted as the same every day for thirty days retrospectively. Since the USD/TL exchange rate data is announced by the TCMB every weekday, the daily data is used without any changes. The financial data of the companies are published quarterly, and the announced quarterly data has been retrospectively accepted as the same every day for ninety days. All the data studied within the scope of the article are for the time interval of 01.10.2016 – 30.09.2021 and weekday data were used. The characteristics and explanations of the data used in the study are given in Table 2. The data used are 10 types for each day and include values for a total of 1256 days.

Table 2 Features used for stock analysis

Feature Name	**Description**
Date	Transaction day
Closing Value (TL)	The closing price of the stock
Volume (TL)	Multiplying the total amount of trading in the relevant stock during the day with the price of the stock at the time of the transaction.
Market Value (mn TL)	The value found by multiplying the share price of the company with the total number of shares
Current Rate	The ratio of current assets of the company to short-term liabilities
Acid Test Ratio	The ratio of the value resulting from the deduction of company stocks from current assets of the company to short-term liabilities
Cash Ratio	Ratio of cash and cash equivalents of the company to short-term liabilities
Net Profit Margin	The ratio of the company's net profit to its net sales
Dollar exchange rate	USD/TL rates announced by the TCMB every weekday
Interest rates	Interest rates announced weekly by the TCMB
Inflation Value	Inflation rates announced monthly by TUIK

Some operations were performed to select the best features to use in the prediction model. First, the 'Date' variable has been removed since the period that the data represents is now known. Data for all features and all companies were examined in detail, and missing value and outlier values were checked. Values that are far outside the general limits of the data are considered outliers. Outlier analysis on our data set in the proposed study was evaluated with the box plot approach, and some results are given in Figure 2. As a result of the examinations made on both the features in the data set and the figures, it has been determined that all the features are numerical and there are no missing or outlier values.

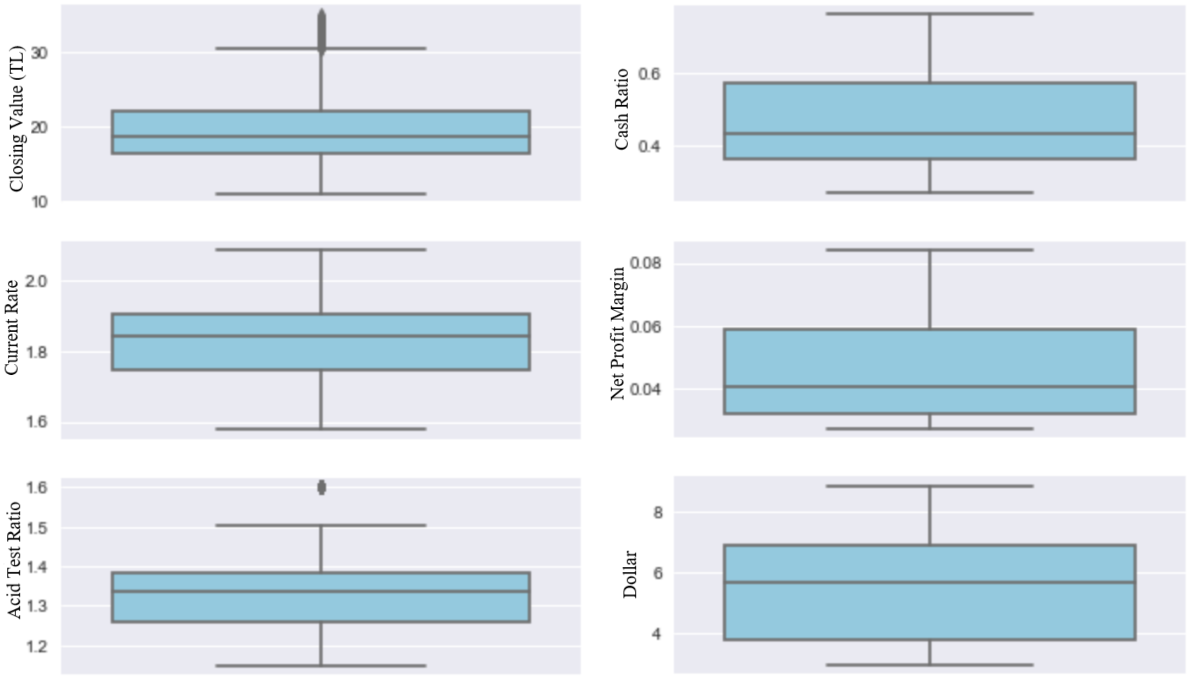

Figure 2 Outlier analysis with boxplot of the features

In order to prevent the machine learning model from being exposed to the multicollinearity problem caused by similar features, correlation analysis between features was performed. The results of the correlation analysis are given in Figure 3. Variables with more than 80% correlation in Figure 3 were excluded from the analysis in order not to cause multicollinearity problems. According to Figure 3, there is a 99% correlation between the closing prices and the Market value. There is a 91% correlation between the current rate and the Acid test rate. There is an 87% correlation between the inflation rate and the interest rates. Market value, Acid test rate and Inflation rate data were removed from the data to eliminate the multicollinearity problem.

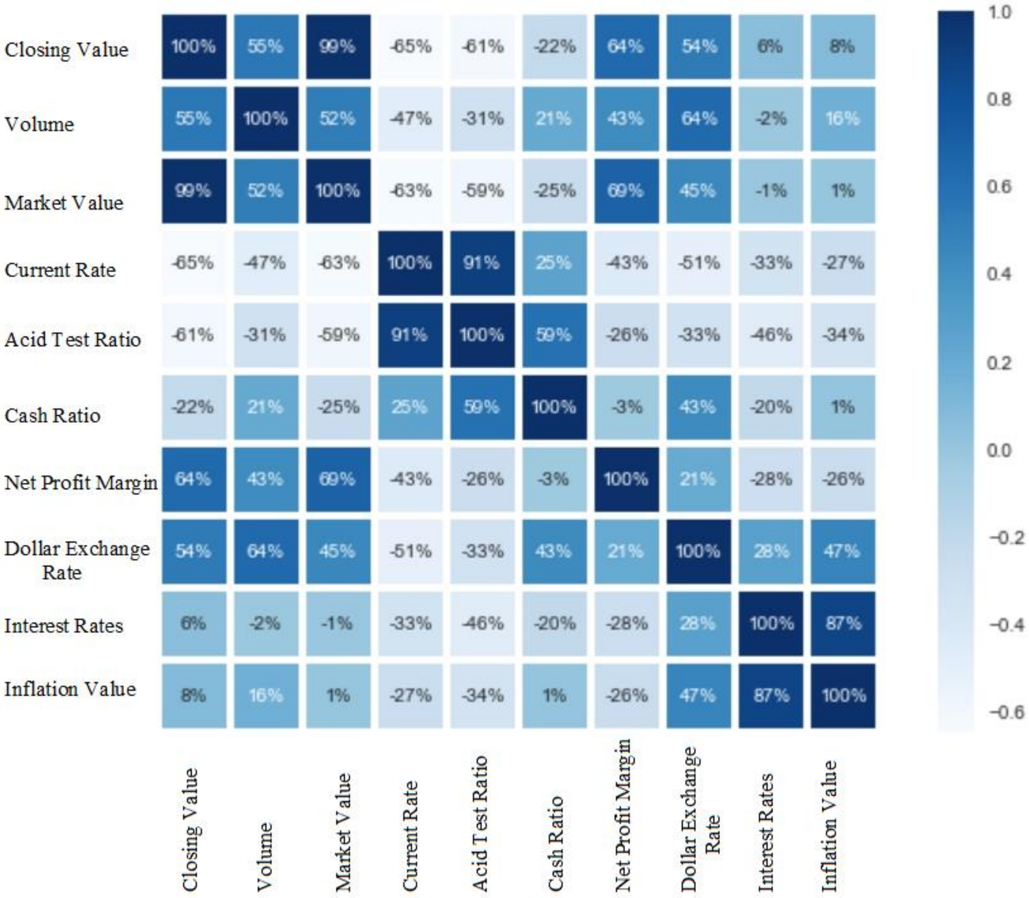

Figure 3 Correlation analysis results of the features

To obtain the experimental results in the proposed study, all algorithms are coded in Python language. The data obtained were separated as training and test data, and it was tried to determine which of them was more successful in which sector by using various algorithms. For the prediction model to classify the data it will see for the first time more successfully, 5-fold cross validation is used on the training set.

Table 3 The hyperparameter space for the machine learning models

Model	Hyperparameter Space
SVM	C: [0.1, 0.5, 1, 1.5] epsilon: [0.01, 0.1, 1] gamma: ['auto'] kernel: ['linear', 'poly', 'rbf'] degree: [2,3,5]
KNN	n_neighbors: [4-20] p: [1,2,3]
XGBoost	learning_rate: [0.01, 0.02, 0.09] The maximum depth: [2, 3, 4, 5, 6] Number of estimators: [100, 200, 500, 2000]
RF	The maximum depth: [80, 90, 100, 110] max_features: [2, 3] min_samples_leaf: [3, 4, 5] min_samples_split: [8, 10, 12] Number of estimators:[100, 200, 300, 1000]

In this approach, the training set is divided into 5 equal parts, each time 4 of these parts are used for model training and one for validation [25][26]. In this study, the dataset is divided into two parts as 70% training and validation set, and 30% test set. After the training process was over, the final success of the model was tested on the test set that aside and never encountered in the training process. SVM, KNN, XGBoost and RF regression models were used to predict future prices of stocks. The performance of regression models is directly related to the hyperparameters selected for the model to run. For the optimization of algorithms, it is very important to find the most optimal parameters instead of directly giving values. In this study, to find the optimum parameters, the parameters with the highest accuracy were selected automatically from the hyperparameter space in Table 3 with the grid search approach.

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|y_k - \hat{y}_k| \tag{1}$$

$$MAPE = \frac{1}{N}\sum_{k=1}^{N}\frac{|y_k - \hat{y}_k|}{\max(\varepsilon, |y_k|)} \tag{2}$$

$$MSE = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2 \tag{3}$$

$$R^2 = 1 - \frac{\sum_{k=1}^{N}|y_k - \hat{y}_k|}{\sum_{k=1}^{N}|y_k - \bar{y}_k|} \tag{4}$$

$$\bar{y}_k = \frac{1}{N}\sum_{k=1}^{N} y_k \tag{5}$$

For the performance measurement of the experimental results, the real values of the stocks and the values predicted by the model were compared. Mean absolute error (MAE), mean absolute percent error (MAPE), mean square error (MSE), and R^2 metrics were used for comparisons. y_k k. actual value, $\hat{y}_k$ k. The metrics used to be the predicted value are given in Equation 1 to Equation 4 for test set size N. A small positive number ϵ is defined to prevent the value of the MAPE score from going to infinity where y_k is zero.

A lower value in the MAE, MAPE, and MSE metrics indicates a better estimate. The value of the R^2 score, which is the most popular metric in linear regression models, is usually between 0 and 1, although sometimes it can be negative. At the highest regression fit, the value of the R^2 score approaches 1. The $\bar{y}_k$ used in calculating R^2 represents the average of all real values as shown in Equation 5.

Table 4 MAPE results for all models of stock price prediction algorithms

	Prediction Models			
Stock Name	**SVM**	**KNN**	**XGBoost**	**RF**
SAHOL	0,035	0,027	0,023	0,023
KCHOL	0,052	0,031	0,023	0,025
EREGL	0,054	0,035	0,029	0,032
KRDMR	0,061	0,041	0,037	0,043
TUPRS	0,078	0,028	0,025	0,026
PETKM	0,047	0,029	0,024	0,027
ARCLK	0,053	0,027	0,022	0,024
VESTL	0,084	0,045	0,043	0,044
TRCELL	0,037	0,024	0,022	0,023
TTKOM	0,046	0,031	0,024	0,028
THYAO	0,081	0,046	0,028	0,031
PGSUS	0,155	0,049	0,046	0,047
Average	**0,065**	**0,034**	**0,029**	**0,031**

The experimental results obtained in the tests carried out based on companies and sectors are given below. First, MAPE error values for all forecasting models are given in Table 4. A low MAPE value indicates that the best estimation is made with the relevant regression model. According to Table 4, when the average MAPE values were examined, the lowest average was obtained as 0.029 for XGBoost. The second-best performance was obtained with RF as 0.031, while the lowest estimation results were obtained with SVM as 0.065. Although the estimation values on a share basis are close to each other, the shares of TRCELL and ARCLK companies reached the best estimation with MAPE values of 0.022.

Secondly, in the experimental results, all prediction models were evaluated in terms of the R^2 metric. In Table 5, the values of the R^2 metric obtained by estimating the shares for four different models are given. The high value of R^2 indicates that the best estimation is made with the relevant regression model. When the average R^2 values are examined according to Table 5, the best value was obtained as 0.989 for XGBoost. The second-best performance was obtained with RF as 0.987, while the lowest predictive value was obtained with SVM as 0.946. While the shares of seven companies have R^2 values of 0.990 and above, the shares of EREGL, KRDMR and VESTL have reached R^2 values of 0.995 and above. When the results in Table 5 are analyzed by sector, it is seen that the models used make the best estimation for the iron and steel sector and the white goods sector.

Table 5 R^2 results for all models of stock price prediction algorithms

	Prediction Models			
Stock Name	**SVM**	**KNN**	**XGBoost**	**RF**
SAHOL	0,920	0,949	0,965	0,966
KCHOL	0,902	0,961	0,981	0,976
EREGL	0,987	0,992	0,996	0,995
KRDMR	0,987	0,991	0,995	0,993
TUPRS	0,888	0,985	0,988	0,988
PETKM	0,971	0,985	0,993	0,991
ARCLK	0,970	0,990	0,994	0,993
VESTL	0,978	0,993	0,995	0,994
TRCELL	0,968	0,985	0,989	0,988
TTKOM	0,951	0,976	0,988	0,983
THYAO	0,911	0,972	0,990	0,987
PGSUS	0,917	0,986	0,992	0,991
Average	**0,946**	**0,980**	**0,989**	**0,987**

Among the forecasting models, the best results were obtained with the XGBoost model, both on a sectoral and company basis. Therefore, XGBoost has come to the fore as the most successful model. In Table 6, the estimation results of all companies' stocks with the XGBoost model are evaluated in terms of MAE, MAPE, MSE and R^2 metrics. MAE, MAPE and MSE values are preferred to be low as they indicate estimation errors. According to Table 6, MAE and MSE values of TUPRS and PGSUS stocks are quite high compared to other stocks.

Tablo 6 Results of forecasting stock prices with the XGBoost model for all metrics

	Metrics			
Stock Name	**MAE**	**MAPE**	**MSE**	**R^2**
SAHOL	0,190	0,023	0,069	0,965
KCHOL	0,366	0,023	0,270	0,981
EREGL	0,217	0,029	0,113	0,996
KRDMR	0,116	0,037	0,032	0,995
TUPRS	2,469	0,025	11,153	0,988
PETKM	0,086	0,024	0,016	0,993
ARCLK	0,424	0,022	0,352	0,994
VESTL	0,386	0,043	0,314	0,995
TRCELL	0,242	0,022	0,106	0,989
TTKOM	0,133	0,024	0,032	0,988
THYAO	0,341	0,028	0,233	0,990
PGSUS	1,722	0,046	6,996	0,992
Average	**0,558**	**0,029**	**1,640**	**0,989**

The lowest MAE values were obtained for PETKM, KRDMR and TTKOM shares as 0.086, 0.116 and 0.133, respectively. The lowest MSE values were also 0.016, 0.032 and 0.032 for the same shares, respectively. Considering all metrics and evaluated on a share basis, the best estimation is made in KRDMR shares. According to error metrics, the worst estimation is for PGSUS stocks. For the sector-based evaluation, the best estimations were made for the companies in the Iron-Steel and Petroleum field.

One of the most important contributions of the proposed study is to examine the effect of current macro changes on the forecasting model. As an example, the effect of the changes in the Central Bank Governors, which took place three times in the 5-year period, on the forecast was investigated. To include this variable in the model, it is assumed that the effect on the market will continue for three months from the date of the change of Governors. A new synthetic variable is used in addition to the previous features for the relevant 90-day period. The results after adding this variable to the model as an independent variable are shown in Table 7. As stated above, since the most successful model is XGBoost, this model has been compared. For all sectors, the effect on share prices before and after the change of Governor was evaluated over the R^2 value.

As can be seen from the table, a decrease is observed in the success of stock forecasting in most companies. While the forecast success of PETKM and ARCLK stocks increased insignificantly, success decreased in SAHOL, KCHOL and TUPRS stocks. This is because the unpredictable impact of the chairman change on policies also affects forecast performance. From an economic point of view, it is thought that the information of companies with foreign exchange deficit and surplus will be important. Looking at the markets after the Central Bank chairman changes, it is seen that there was an increase in the dollar exchange rate and a decrease in stock prices.

Table 7 The effect of the Central Bank's governors changes on the stock price prediction

Stock Name	**R^2 Before Governor Change**	**R^2 After Governor Change**	**Difference**
SAHOL	0,9650	0,8990	-0,0660
KCHOL	0,9810	0,9345	-0,0465
EREGL	0,9960	0,9926	-0,0034
KRDMR	0,9950	0,9793	-0,0157
TUPRS	0,9880	0,9057	-0,0823
PETKM	0,9930	0,9951	0,0021
ARCLK	0,9939	0,9943	0,0005
VESTL	0,9953	0,9864	-0,0089
TRCELL	0,9892	0,9901	0,0009
TTKOM	0,9880	0,9851	-0,0030
THYAO	0,9905	0,9682	-0,0223
PGSUS	0,9922	0,9875	-0,0047

Although the increase in the dollar exchange rate seems to be a positive development in terms of balance sheet for companies with excess foreign exchange position, these stocks also decline with the effect of the general decline in the stock market index, and thus, the explanation success of the model decreases due to the share price, which is negatively affected because of the positive development. For companies with short foreign exchange position, the increase in the dollar exchange rate is a negative development in terms of the balance sheet, and these stocks also decline with the effect of the general decline in the stock market index, and thus, the share price is adversely affected because of the negative development.

4. Conclusions

In this study, the prices of the stocks of companies in different sectors traded in the BIST 30 Index were examined. The data of two companies selected as examples from Holding, White Goods, Petrochemical, Iron and Steel, Air Transport and Communication sectors were examined. The study has shown that the Xgboost algorithm is the most successful algorithm based on both sectors and companies. While the RF algorithm has the second-best performance on the basis of industry and company, the worst performance belongs to the SVM algorithm. In the current literature studies, BIST30 and BIST100 index estimations are made. When stock-based forecasting was made, stocks were randomly selected. In the proposed study of this article, the effects of some specific periods such as the fact that a sector-based analysis was made and the change of the central bank governor effect on the sectors were examined by machine learning methods. As a result of the studies, it has been determined that the machine learning-based estimation of stocks in the iron and steel, petrochemical and communication sectors has achieved more successful results.

Acknowledgments

This study was partially carried out in the Software Technologies Research Laboratory (STAR Lab) of the Kocaeli University Software Engineering Department.

References

[1] I. K. Nti, A. F. Adekoya and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," *Artificial Intelligence Review*, 53(4), pp. 3007-3057, 2020.

[2] H. Dağlı, "Sermaye Piyasası ve Portföy Analizi," 3rd Ed., *Derya Kitabevi*, Trabzon, 2009.

[3] S. Tekin, "Destek vektör makineleri yöntemi ile İMKB 100 endeksi hareket yönü tahmini" *Uşak University Social Sciences Institute*, Master Thesis, Uşak, 2013.

[4] U Demirel, "Hisse senedi fiyatlarının makine öğrenmesi yöntemleri ve derin öğrenme algoritmaları ile tahmini", *Giresun University Social Sciences Institute*, Master Thesis, 2019

[5] P. Chhajer, M. Shah and A. Kshirsagar, "The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction," *Decision Analytics Journal*, 2, 100015, 2022.

[6] Z. D. Akşehir and E. Kılıç, "Prediction of Bank Stocks Price with Machine Learning Techniques", *TBV Journal of Computer Science and Engineering,* 12 (2) , pp. 30-39, 2019.

[7] E. Filiz, H. A. Karaboğa and S. Akoğul, "Bist-50 index change values classification using machine learning methods and artificial neural networks", *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 26(1), pp. 231-241, 2017.

[8] H. S. Sim, H. I. Kim and J. J. Ahn, "Is Deep Learning for Image Recognition Applicable to Stock Market Prediction", Complexity, 4324878, 2019.

[9] Z. Ivanovski, N. Ivanovska and Z. Narasanov, "The regression analysis of stock returns at MSE", *Journal of Modern Accounting and Auditing*, 12(4), pp. 217-224, 2016.

[10] G. Şişmanoğlu, F. Koçer, M. Önde and O. K. Sahingoz, " Price Forecasting in Stock Exchange with Deep Learning Methods ", *BEU Journal of Science*, 9(1), pp. 434-445, 2020.

[11] V. Gururaj, V.R. Shriya, and K. Ashwini, "Stock market prediction using linear regression and support vector machines", *Int J Appl Eng Res*, 14(8), 1931-1934, 2019.

[12] S. Karasu, A. Altan, S. Bekiros and W. Ahmad, "A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series", *Energy*, 212, 118750, 2020.

[13] N. K. Ustalı, N. Tosun, and Ö. Tosun, "Stock Price Forecasting Using Machine Learning Techniques", *Eskişehir Osmangazi University Journal of Economics and Administrative Sciences*, 16(1), pp. 1-16, 2021.

[14] M. E. Arslan and P. Kırcı, "Stock Market Analysis with Machine Learning". *European Journal of Science and Technology*, (28), pp. 1117-1120, 2021.

[15] S. Arslankaya and Ş. Toprak, "Using Machine Learning and Deep Learning Algorithms for Stock Price Prediction", *International Journal of Engineering Research and Development*, 13(1), 178-192, 2021.

[16] Y. C. Chen and W. C. Huang, "Constructing a stock-price forecast CNN model with gold and crude oil indicators", *Applied Soft Computing*, 112, 107760, 2021.

[17] Z. D. Akşehir and E. Kılıç, "A new rule-based approach for encountered data imbalance problem in stock predicition and 2D-CNN model", *TBV Journal of Computer Science and Engineering*, 15 (1), pp. 6-13, 2022.

[18] M. Leippold, Q. Wang and W. Zhou, "Machine learning in the Chinese stock market", *Journal of Financial Economics*, 145(2), pp. 64-82, 2022.

[19] V. V. Prasad, S. Gumparthi, L.Y. Venkataramana, S. Srinethe, R. M. Sruthi Sree and K. Nishanthi, "Prediction of Stock Prices Using Statistical and Machine Learning Models: A Comparative Analysis", *The Computer Journal*, 65(5), 1338-1351, 2022.

[20] V. Vapnik, S. Golowich and A. "Smola, Support vector method for function approximation, regression estimation and signal processing", *Advances in neural information processing systems*, 9, 1996.

[21] S.B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background", *Internat. J. Eng. Res. Appl*. 3(5), 605-610, 2013.

[22] L. Huang, Y. Li, S. Chen, Q. Zhang, Y. Song, J. Zhang and M. Wang, "Building safety monitoring based on extreme gradient boosting in distributed optical fiber sensing", *Optical Fiber Technol.*, 55, 102149, 2020.

[23] S. Obata, C. J. Cieszewski, R. C. Lowe III, and P. Bettinger, "Random Forest Regression Model for Estimation of the Growing Stock Volumes in Georgia, USA, Using Dense Landsat Time Series and FIA Dataset" *Remote Sensing*, 13(2), 218.

[24] IS Investment, , “Market Data,” 2022. [Online]. Available: https://www.isyatirim.com.tr. [Accessed: 24-May-2022].

[25] F. Alareqi and M. Z. Konyar , "High Accuracy Classification of Covid-19 from CT Images Using Transfer Learning Architectures", *Dicle University Journal of Engineering*, 13(3), pp. 457-466, 2022

[26] F. Al-Areqi and M. Z. Konyar, "Effectiveness evaluation of different feature extraction methods for classification of covid-19 from computed tomography images: A high accuracy classification study", *Biomedical Signal Processing and Control*, 76, 103662, 2022.