applied sciences

MDPI

Article

Predicting Financial Performance in the IT Industry with Machine Learning: ROA and ROE Analysis

Burçin Tutcu [1,*], Mehmet Kayakuş [2], Mustafa Terzioğlu [1], Güler Ferhan Ünal Uyar [3], Hasan Talaş [1,*] and Filiz Yetiz [4]

[1] Accounting and Tax Department, Korkuteli Vocational School, Akdeniz University, Antalya 07800, Türkiye; mterzioglu@akdeniz.edu.tr
[2] Department of Management Information Systems, Faculty of Social and Human Sciences, Akdeniz University, Antalya 07800, Türkiye; mehmetkayakus@akdeniz.edu.tr
[3] Department of Business Administration, Faculty of Economics and Administrative Sciences, Akdeniz University, Antalya 07058, Türkiye; guleruyar@akdeniz.edu.tr
[4] Department of Finance and Banking, Faculty of Applied Sciences, Akdeniz University, Antalya 07058, Türkiye; filizyetiz@akdeniz.edu.tr
* Correspondence: burcintutcu@akdeniz.edu.tr (B.T.); htalas@akdeniz.edu.tr (H.T.)

Abstract: IT is recognized as the engine of the digital world. The fact that this technology has multiple sub-sectors makes it the driving force of the economy. With these characteristics, the sector is becoming the center of attention of investors. Considering that investors prioritize profitability, it becomes a top priority for managers to make accurate and reliable profitability forecasts. The aim of this study is to estimate the profitability of IT sector firms traded in Borsa Istanbul using machine learning methods. In this study, the financial data of 13 technology firms listed in the Borsa Istanbul Technology index and operating between March 2000 and December 2023 were used. Return on assets (ROA) and return on equity (ROE) were estimated using machine learning methods such as neural networks, multiple linear regression and decision tree regression. The results obtained reveal that the performance of artificial neural networks (ANN) and multiple linear regression (MLR) are particularly effective.

Keywords: ROA; ROE; machine learning; financial forecasting

check for updates

Citation: Tutcu, B.; Kayakuş, M.; Terzioğlu, M.; Ünal Uyar, G.F.; Talaş, H.; Yetiz, F. Predicting Financial Performance in the IT Industry with Machine Learning: ROA and ROE Analysis. *Appl. Sci.* **2024**, *14*, 7459. https://doi.org/10.3390/app14177459

Academic Editor: Douglas O'Shaughnessy

Received: 19 July 2024
Revised: 21 August 2024
Accepted: 21 August 2024
Published: 23 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the new world order, where digitalization is at the center, it is known that the only way to ensure sustainable growth is through technology breakthroughs. Some of the world's leading economists describe the current period in the economy as the "IT (Information Communication Technologies) Techno-Economic Paradigm". In this period, which is still in its early stages, it is expected that the activities of the IT sector, such as hardware, software, communication and telecommunication services, and content production of information technology services, are both the driving force behind the economy and can radically change activities in other sectors of the economy. The global IT market size will grow by 13% to USD 4.3 trillion in 2023, while the information technology market size will increase by 25.7% and the communication technology market size will increase by 1.4%, as reported by TÜBİSAD. In 2026, this increase is projected to reach USD 5.6 trillion.

The IT sector has consistently been an attractive sector for investors since the mid-1990s. However, after the bursting of the dot-com bubble in 2000, it is seen that the profitability of these companies is prioritized in the investment feasibility of these companies, not the idea or product that the company comes up with. In the same way, the managers of these companies prioritize the realized profitability and aim to increase the profitability rates of the companies in order to support investors' decisions.

In this process, ROA (return on assets) and ROE (return on equity) are the ratios that managers especially work on. ROA shows how much profit a company makes for the assets it owns, and it shows how efficiently it uses its assets. ROE, on the other hand, expresses how effectively the investments made by shareholders in the company are used to generate profit. In this respect, it is known to be one of the most important profitability indicators for investors. The predictability of ROA and ROE plays an important role in the growth and sustainability of companies. Accurate estimation of these indicators, on the other hand, will form the basis for strategic decisions to be taken by managers.

In profitability forecasting, machine learning methods are known to provide more successful results in different sectors than traditional methods such as regression. The reason for this is that these methods generally produce effective solutions in cases of nonlinearity and multicollinearity in financial data sets.

The fact that the IT sector is considered as the locomotive of the sustainable world makes the decisions to be taken by managers much more important. The fact that managers use past period data in their decisions makes future forecasts much more important. The motivation of this study is to increase the number of investors by predicting profitability in the IT sector, which is the key to the digital world, and to determine the sustainability of businesses. For this purpose, profitability studies conducted with machine learning methods in different sectors will first be examined in the literature section. Then, the model designed in the methodology section will be tested using machine learning methods, and the results will be discussed.

This study differs from other studies in that it examines the IT sector, which has not been addressed as a sector in the literature before. The fact that companies in the IT sector need to follow the pace of technology and make extra efforts in innovation compared to companies in other sectors requires these companies to have a strong financial structure and the potential to generate profit. Therefore, the study will contribute to the literature in terms of creating the ideal model with ideal machine learning methods.

2. Literature

It is known that ROA and ROE are among the indicators used to measure, evaluate and analyze the performance of enterprises. When the literature is examined, it is seen that these ratios provide important information about the profitability, performance and financial structure of the enterprise. ROA is used to measure how efficiently a company uses its assets, while ROE provides information about the efficiency of the use of the company's equity. Accurate and reliable analysis of these ratios is critical for both internal and external users of information.

When the studies on the importance of ROA in the literature are examined, it is seen that these studies are especially on determining how effectively the assets of the enterprise are used. For example, Penman (2013) and Damodaran (2007) emphasize in their study that a high ROA indicates that the enterprise earns high profits by using its assets efficiently, and this shows the operational efficiency and management success of the enterprise. In this context, it is known that ROA is effective in all operational processes of the enterprise, from production processes to marketing activities [1,2].

ROE is considered in the literature as a critical tool for assessing companies' return on equity and shareholder returns. In their study, Brigham and Ehrhardt (2017) emphasize that a high ROE indicates that the firm generates high returns by using its equity effectively, which is attractive to investors. At the same time, ROE is also used to assess the effectiveness of a firm's financial strategies and capital structure. There are many studies in the literature that ROE also reflects the sustainable growth rate and competitiveness of the business [3].

The relationship between ROA and ROE and the evaluation of these ratios together are frequently emphasized in the literature. There are many studies on this subject. For example, Higgins (2012) and Ross, and Westerfield and Jaffe (2016) state that analyzing these two ratios together allows for a more holistic understanding of business performance. ROA focuses on the efficiency of assets, while ROE focuses on return on equity. Analyzing the

two ratios together provides information about a company's debt utilization and financial risk management. For this reason, ROA and ROE have been included in the literature as key financial indicators that should be considered together for a comprehensive evaluation of business performance. Table 1 presents studies that include ROA and ROE [4,5].

Table 1. Studies on ROA and ROE in the literature.

Author/Year	Article Title	Variables	Machine Learning Technique
Pham Vu Hong Son, Le Tung Duong, 2024 [6]	Research on applying machine learning models to predict and assess return on assets (ROA)	Days sales outstanding, Days inventory outstanding, Days payable outstanding, Cash conversion cycle, Networking capital, Financial leverage, Current ratio, Growth rate, Return on equity, Earnings before interest taxes, Gross margin	Lasso, Ridge, K Neighbor Regressor, SVR, Random Forest, GBR, XGBoost, MLP Regression
Caterina De Lucia, Pasquale Pazienza, and Mark Bartlett, 2020 [7]	Does Good ESG Lead to Better Financial Performances by Firms? Machine Learning and Logistic Regression Models of Public Enterprises in Europe	Emission score, CO_2 equivalent emissions, Resource use score, ESG score, Environmental innovation score, Salary gap, Number of employees in the CSR reporting, Number of women employees, Workforce score, Management score, Operating income, Net income, Change in total equity	Random forest, Support Vector regression, K-Nearest Neighbor Model, Artificial Neural Network Model, Ridge Regression
Ting-Hsuan Chen, Rong-Cih Chang, 2021 [8]	Using machine learning to evaluate the influence of FinTech patents: The case of Taiwan's financial industry	(Interest income − Interest expense)/Total assets, Net income after tax/Operating income, Log(Total assets), Qualified capital/Risk weighted assets, Patents	Logistic Regression, Multilayer Perceptron, Decision Tree, Random Forest, Naive Bayes Classifier, Bayesian Network, Support Vector Machine
Pratibha Rai, Bibhuti Bhusan Mohapatra, Vanita Jain, 2023 [9]	Major Determinants of Bank Profitability in India: A Machine Learning Approach	Financial crisis, Unemployment, Inflation, GDP growth rate, Real effective exchange rate	The random forest (RF) classification algorithm is executed using the CARET package in R. The results obtained from feature selection are corroborated the RF classification findings.
Potta Chakri, Saurabh Pratap, Lakshay, Sanjeeb Kumar Gouda 2023 [10]	An exploratory data analysis approach for analyzing financial accounting data using machine learning	Total Revenue	Linear Regression, KNN (K Nearest Neighbor), SVR (Support Vector Regression), and Decision Tree
Stewart Jones, William J. Moser, Matthew M. Wieland 2023 [11]	Machine learning and the prediction of changes in profitability	On Average Rnoa, Net Operating Assets	TreeNet, OLS Regression, PZ Regression
Nicholas D. Belesis, Georgios A. Papanastasopoulos and Antonios M. Vasilatos 2023 [12]	Predicting the Profitability of Directional Changes Using Machine Learning: Evidence from European Countries	ROA/ROE/RNOA/FCF/CFO	Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), Decision Trees (DT), Random Forest (RF)
Xinqi Dong, Bo Dang, Hengyi Zang, Shaojie Li, Danqing M 2024 [13]	The Prediction Trend of Enterprise Financial Risk based on Machine Learning ARIMA Model	Profit, Revenue, Expenditure, Assets, Liabilities	Support Vector Machine, Logistic Regression, Neural Network, ARIMA regression model
Chuqing Zhang, Han Zhang, AND Dunnan Liu 2019 [14]	A Contrastive Study of Machine Learning on Energy Firm Value Prediction	Deal value Input: EBIT, ROE, ROA, CAPEX, M&A Type, Asset Turnover, Cash Debt Ratio, Total Debt to Assets, Firm Type, Nationality, Acquisition Year, Share	Decision Tree Regression, Supported Vector Regression, Artificial Neural Network
Andrey Zahariev, Petko Angelov, Silvia Zarkova (2022) [15]	Estimation of Bank Profitability Using Vector Error Correction Model and Support Vector Regression	ROE and ROA	Support Vector Regression and Vector Error Correction Model
Kristof Tamas and Miklos Virag (2022) [16]	What drives financial competitiveness of industrial sectors in Visegrad Four countries? Evidence by use of machine learning techniques.	ROA and ROE	The K-Nearest Neighbor (KNN) and Random Forest (RF) Methods

Table 1. *Cont.*

Author/Year	Article Title	Variables	Machine Learning Technique
Vic Anand, Robert Brunner, Kelechi Ikegwu, Theodore Sougiannis (2019) [17]	Predicting Profitability Using Machine Learning	ROA (return on assets), ROE (return of equity), CFO (cash flow from operations), RNOA (return on net operating assets) and FCF (free cash flow)	Random Forests
Md. Nur-Al-Ahad, Syeda Nusrat, Vagavi Prakash (2019) [18]	Nexus Between Corporate Governance and Firm Performance in Malaysia: Supervised Machine Learning Approach	ROA and ROE	Supervised Machine Learning Approach

In the literature, it is known that machine learning methods give successful results in profitability forecasting. However, it has been determined that ROA and ROE analysis has not been conducted in the IT sector, which is the key to the future. For this reason, this study is expected to contribute to the literature.

3. Materials and Methods

In this study, three machine learning methods were used to predict ROA and ROE values. Artificial neural networks, multiple linear regression and decision tree regression methods were used for forecasting. The choice of methods such as neural networks, multivariate linear regression, and decision tree regression in ROA and ROE forecasting offers various advantages depending on the modeling needs. Neural networks provide superiority in modeling complex relationships; multivariate linear regression offers simple and interpretable relationships; and decision trees are effective in understanding heterogeneous and non-linear data structures. In selecting each method, factors such as the nature of the data set, prediction accuracy requirements, and model understandability were taken into account. The model consists of 14 independent variables and 1 dependent variable.

Neural networks are primarily chosen for their ability to model complex and non-linear relationships. These methods have the capacity to learn complex relationships between data and can provide a high degree of abstraction and feature learning due to their multilayer structure. Neural networks can exhibit strong performance, especially in large data sets and with many variables. In the prediction of financial indicators such as ROA and ROE, neural networks can model the non-linear structure of the relationships of these indicators with economic and financial factors. Moreover, the performance of the model can be improved regularly, and more accurate predictions can be made by hyperparameter optimization.

Multivariate linear regression is a basic method for modeling linear relationships between a dependent variable and multiple independent variables. This approach may be suitable for modeling linear relationships that are frequently observed in financial data. In the estimation of financial ratios such as ROA and ROE, this method is used to analyze linear relationships between independent variables. The advantages of multivariate linear regression include the understandability of the model and the interpretability of the results. In addition, the model is relatively simple to set up and implement and does not require much data.

Decision tree regression is used to model the relationships between different groups and structures in a data set and makes predictions by dividing the data into various decision points. This method can be particularly effective when there are significant thresholds and non-linear relationships between data points. Financial indicators such as ROA and ROE may show different effects under different economic and financial conditions, so decision trees can be useful for analyzing such heterogeneous data structures. Decision trees clearly visualize the interactions of variables and important features within the data, which facilitates the interpretation of the model. Furthermore, this method can be more robust to noise and missing values in the data set.

3.1. Data Set

The data in the study belonged to 13 technology firms included in the Borsa Istanbul Technology index and operating between March 2000 and December 2023. Table 2 shows the companies whose data were used in the study.

Table 2. Technology firms in the Borsa Istanbul Technology Index.

Companies	Codes
Alcatel Lucent Teletas Telecommunications Co. (Boulogne-Billancourt, France)	ALCTL
Arena Computer Industry and Trade Co. (İstanbul, Turkey)	ARENA
Despec Computer Marketing and Trading Co. (Dubai, United Arab Emirates)	DESPC
Datagate Computer Materials Trading Co. (Nicosia, Cyprus)	DGATE
Fonet Information Technologies Co. (Ankara, Turkey)	FONET
Indeks Computer Systems Engineering Industry and Trade Co. (İstanbul, Turkey)	INDES
Ingram Micro Computer Systems Co. (Irvine, CA, USA)	INGRM
Karel Electronics Industry and Trade Co. (Istanbul, Turkey)	KAREL
Kron Technology Co. (Jersey City, NJ, USA)	KRONT
Link Computer Systems Software and Hardware Industry and Trade Co. (Istanbul, Turkey)	LINK
Logo Software Industry and Trade Co. (Gebze, Turkey)	LOGO
Netas Telecommunications Co. (Istanbul, Turkey)	NETAS
Plastikkart Smart Card Communication Systems Industry and Trade Co. (Istanbul, Turkey)	PKART

The data of these companies were obtained from the Finnet financial analysis program and edited by the researchers. The data range of the study covers the period between March 2000 and December 2023. Quarterly financial period data of the companies were used as data frequency. In this study, there were 27 fiscal periods for each company. A total of 810 data points for each company and 10,530 data points in total were analyzed. Table 3 presents the variables of the study.

In the business finance literature, financial ratios consist of five main groups. These are liquidity ratios, financial structure ratios, profitability ratios, operating ratios and stock market ratios. With these groups of ratios, the situation of the company in general and the sector in particular is taken into consideration, and the model is created through dependent variables. Vic Anand et al. (2019), Thin-Hsuan, Chen and Rong-Cih Chang (2021), Stewart, J et al. (2023), Phan Vu Hang and Le Tung Duong (2024), and Xinqi Dung et al. (2024) created models in the same way to predict ROA and ROE in their studies. While designing the model of the study, the main variables related to ROA and ROE, which are both used in corporate finance and have an important place in the IT sector, were selected as independent variables. Especially for the IT sector, operating ratios were emphasized. This is because it is considered important to manage the operational processes of the sector. The designed model was applied to each company separately using artificial neural networks (ANN), multiple linear regression (MLR) and decision tree regression (DTR) methods [6,8,11,13,17].

A preliminary analysis was performed to assess whether the data were balanced and the distribution of the data. For this, it was checked whether there were missing values in the data. Statistical methods were used to determine whether there were outliers in the data. The data set was checked for any inconsistencies, incorrect entries, or data format errors. The central tendency of the distribution was determined by calculating the mean and median values of each variable. Standard deviation and variance were calculated to examine the spread of the data set. This shows how spread out variables are. This preliminary analysis provided a basic framework for understanding whether the data were balanced and the distribution of the data. The findings show that the data do not show a distribution that would have a negative impact on the modeling process.

Table 3. Dependent and independent variables used in the study.

Dependent Variables	
Return of Assets (ROA)	
Return of Equities (ROE)	
Independent Variables	
Receivables Turnover	Operational Ratios
Inventory Turnover	
Fixed Asset Turnover	
Fixed Assets/Total Assets (%)	
Current Assets/Total Assets (%)	
Tangible Fixed Assets Turnover	
Price Earnings Ratio	Stock Exchange Ratios
Earnings per Share	
Cost of Sales/Net Sales (%)	Profitability Ratios
EBITDA Margin (%)	
Net Profit Growth (%)	
Net Profit Margin (%)	
Leverage Ratio (%)	Financial Structure Ratio
Current Ratio	Liquidity Ratio

3.2. Artificial Neural Networks

Artificial neural networks have emerged as a result of the mathematical modeling of the learning process inspired by the human brain. Artificial nerve cells are similar to biological nerve cells. Artificial neurons also form artificial neural networks by establishing connections between them. Just like biological neurons, artificial neurons have sections where they receive input signals, collect and process these signals, and transmit outputs [19].

An artificial neural cell consists of five layers. It consists of input, weights, summation junction, activation function and output layers, as shown in Figure 1.

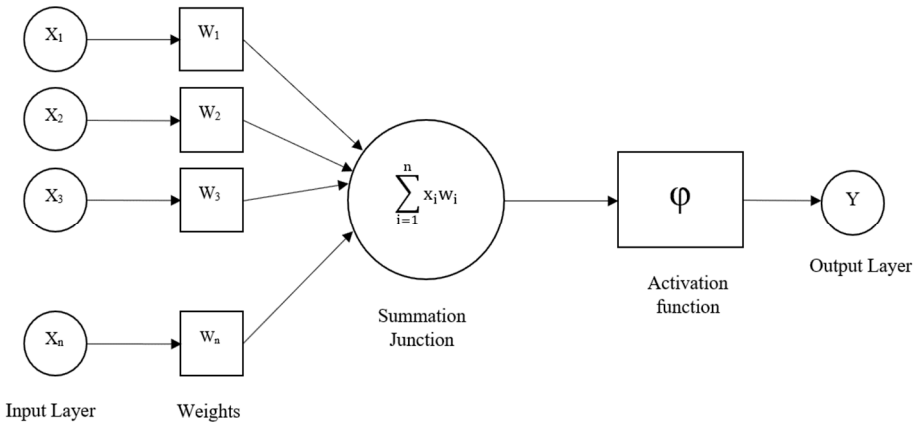


Figure 1. ANN structure.

Inputs to the neural cell can come from other cells or directly from the outside world. The information arriving at the neural cell is multiplied by the weight of the connections they come from before reaching the nucleus through the inputs. In this way, the effect of the inputs on the output to be produced can be adjusted. The values of these weights can be positive, negative or zero. Inputs with a weight of zero have no effect on the output. The

summation function is a function that calculates the net input of a neural cell by adding the inputs multiplied by the weights. For each neuron, the total function is calculated by multiplying the values of the neurons connected to that neuron by the weight.

$$y = \sum_{i=1}^{n} w_i x_i + b \quad (1)$$

Here, y represents the output, x represents the input, w represents the weight, and b represents the bias.

The activation function processes the net input to the cell and determines the output that the cell will produce in response to this input. The value coming out of the activation function is the output value of the cell. This value can either be given to the outside world as the output of the artificial neural network, or it can be reused within the network. The most used activation function today is the "Sigmoid function" shown in Equation (2) [20,21].

$$\varphi(y) = \frac{1}{1 + e^{-y}} \quad (2)$$

Artificial neural networks are divided into feed-forward and feedback networks according to the way the neurons are connected to each other. In feed-forward networks, neurons are in regular layers from input to output. There is only a connection from one layer to the following layers. The information coming into the artificial neural network passes through the input layer and then through the intermediate layers and the output layer, respectively, and then outputs to the outside world. In feedback neural networks, unlike feed-forward neural networks, the output of a cell is not only given as input to the layer of the next cell. It can also be connected as input to any cell in the previous layer or in its own layer [22,23].

3.3. Multiple Linear Regression

Linear regression is a data analysis technique that predicts the value of unknown data using another relevant and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation [24]. The method used to explain the cause–effect relationships between two or more independent variables affecting a variable with a model and to determine the effect levels of these independent variables is called multiple regression analysis [25]. In this respect, it is similar to simple regression. There is more than one prediction variable in the multiple regression equation. Least squares (LS) estimators can be used to estimate the model parameters [26].

The most general multiple regression equation, where X_i are the independent variables and Y is the dependent variable, is as follows

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, 2, 3 \ldots, n \quad (3)$$

In the equation, β_j (j = 1, 2, 3, …, ε_j) represent the regression coefficients, ε_i represent the ith random error values, Y_i represent the ith dependent variable, and X_{ij} (j = 1, 2, 3, …, p) represent the independent variable. The epsilon represents error values, Y is the dependent variable, and X_J are the independent variables.

In multiple linear regression analysis, the contribution of some of the independent variables to the model may be insufficient. For this reason, it is necessary to determine the independent variables that will explain the dependent variable in the "most appropriate" way and to remove unimportant variables from the model [27]. This process is called "variable selection". In multiple linear regression, each independent variable has a different degree of influence on the dependent variable. Therefore, in addition to the equation in simple linear regression, the coefficient of each variable does not have to be the same [28].

3.4. Decision Trees Regression

A decision tree is a flowchart-like diagram that maps all potential solutions to a given problem. Decision trees are in the form of a tree structure that can be built on both regression and classification models [29,30]. Regression is used on numerical target data, while classification is used on categorical data. Tree-based methods are simple and convenient in terms of interpretation.

A decision tree has a predefined target variable. In terms of their structure, they offer a top-down strategy [31]. A decision tree is a structure used to partition a dataset containing a large number of records into smaller sets by applying a set of decision rules, in other words, by applying simple decision-making steps to divide large amounts of records into very small groups of records [32].

The decision tree structure consists of a root node, an internal node and a leaf node. The root node is the top node of the tree. It makes decisions by dividing the dataset between two or more subsets. Interior nodes are the nodes below the root node and above the leaf nodes. They perform a feature test and divide the data into subsets. Leaf nodes are the nodes below the internal nodes. These nodes contain a class label or numeric value, which is the value by which the data point is classified or predicted [33].

One of the most important problems in decision trees is the criteria for partitioning or branching from any root. Examples of algorithms that use entropy-based partitioning are ID3 and its advanced form, C4.5. As for classification and regression trees, Twoing and Gini algorithms can be used [34].

Entropy is a measure of uncertainty. If a dataset has a single label, it has a lower entropy value. Therefore, the data should be divided in a way that minimizes entropy. If the splitting process is good, the predictions are successful.

$$H = -\sum p(x)\log p(x) \tag{4}$$

Information gained is used to determine the best split. Information gain:

$$\text{Gain}(S, D) = H(S) - \sum_{V \in D} \frac{|V|}{|S|} H(V) \tag{5}$$

S is the original data set, and D is a partition of the set. Each V is a subset of S. The information gain is defined as the difference between the entropy of the original dataset before the split and the entropy value of each attribute.

Pruning is the process of removing parts from the decision tree that do not contribute to the classification. This makes the decision tree both simple and understandable. Preliminary pruning is performed when the tree is constructed. If the values of the partitioned attributes are not above a certain threshold (error tolerance), the tree partitioning process is stopped at that point, and the dominant class label in the currently available set is created as a leaf. Subsequent pruning takes place after the tree has been created. This can be performed by deleting subtrees to create leaves, raising subtrees, and cutting branches [35].

4. Results and Discussion

In this study, a series of data preprocessing steps were performed on the data set. The missing data rate was calculated for each variable. If the missing data rate is above 5%, this may affect the data analysis and modeling processes. Missing values can be filled with mean, median or mode. For continuous data, the mean or median is usually used, and for categorical data, the mode is used. For this study, the mean method was used. Since outliers may have a negative impact on the modeling process, these values were removed from the data set. The data were converted to the range [0, 1] with the min–max normalization method. Correlation between features was analyzed for feature selection. The dataset was adapted to the modeling process, and the overall performance of the model was improved.

The forward-elimination method was used for data selection in the study. The forward-elimination method starts from a situation where there are no features in the model and

adds the best-performing feature to the model at each step. This process continues until a certain stopping criterion is reached. These criteria are usually no significant improvement in the performance of the model, a certain number of features, or some other specified target.

Normalization, partitioning and analysis were performed, respectively. In the first stage, the data were normalized according to the specified method. Min–max normalization is a technique frequently used in data preprocessing. This method transforms data expressed in different ranges into a range between 0 and 1, making them comparable. The smallest (minimum) and largest (maximum) values are found for each feature in the data set. For each data point, the minimum value is subtracted from the value for that feature and divided by the difference between the maximum value and the minimum value. The min–max method normalizes the data linearly. While the minimum is the lowest value a data point can take, the maximum is the highest value a data point can take [36].

$$x^i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (6)$$

In this equation, x^i denotes the normalized data; x_i denotes the input value; x_{min} denotes the smallest number in the input set; x_{max} denotes the largest number in the input set.

In this study, R^2 and RMSE statistical metrics were used to analyze the success and error values of the model.

R^2 is a measure of the model's explainability. R^2 is a statistical measure of how close the data are to the fitted regression line and works by measuring the amount of variance in the predictions explained by the dataset. It is also a measure of the difference between the observations in the data set and the predictions generated by the model [37]. R^2 takes a value between 0 and 1. R^2 value of 1 means that the model is perfect, while a value of 0 means that the model performs poorly on a data set that it has not encountered before [38]. The closer R^2 is to 1, the better trained the model is.

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} \quad (7)$$

RMSE measures the average magnitude of the error. RMSE is found by taking the root of the square of the difference between the predicted value and the true value for all values in the data set. RMSE is heavily influenced by outliers and is often used to reconstruct the error distribution [39].

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n} e_j^2}{n}} \quad (8)$$

where e represents the amount of error, and n represents the amount of data.

It is crucial to use a variety of data sets and techniques to verify the applicability and robustness of the model. Training–test separation, cross-validation, external data sets, and comparison with standard data sets are some of these methods. In this study, the data set was divided into two as training and testing. This method allows you to evaluate the performance of the model against data that it has not seen before.

In the training phase, the most appropriate parameters for the success of the model are determined, and in the testing phase, the success of the model is measured according to statistical metrics [40]. Different ratios were tested in this study, and it was decided that the most appropriate ratio was 70% training and 30% testing.

In this study, a feedback neural network model was used for the ANN. Thus, the error was minimized. It was observed that the model with four hidden layers and three neurons in each layer gave the most successful results. The learning coefficient (λ) is a factor that determines the speed of updating the weights in artificial neural networks. While a high value allows the weights to be updated quickly, very large values can lead to instability during the training of the network. The momentum coefficient (α) is a coefficient that increases the effect of previous updates in updating the weights. This helps to prevent the network from becoming stuck at local minima and can speed up the learning process. The

learning coefficient (λ) was chosen to be 0.6, and the momentum coefficient (α) was 0.8. The number of iterations was set as 1000.

In multiple linear regression, there is more than one independent variable in the system. For this reason, the effects of variables on the system may be different from each other. In this study, the forward-elimination method was used for the selection of variables. The model initially contained a single variable, and the number of variables increased with each step.

In the decision tree method, which was another method used in this study, the Gini index technique was used as a quality measure. The Gini index is calculated for each class, and the sum of the squares of the results is subtracted from 1 [41]. The Gini value takes a result between 0 and 1, and the closer the result is to 0, the better the discrimination. If the number of records in the model is less than or equal to 2, the tree is designed to grow no further.

The results of the models using machine learning methods for ROA are presented in Table 4.

Table 4. Estimation results for ROA.

	ANN				MLR				DTR			
	R^2	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE
ALCTL	0.836	0.124	0.025	0.159	0.778	0.114	0.018	0.136	0.683	0.253	0.097	0.312
ARENA	0.707	0.095	0.014	0.119	0.88	0.056	0.007	0.082	0.958	0.258	0.119	0.345
DESPC	0.878	0.077	0.01	0.101	0.774	0.169	0.038	0.196	0.626	0.252	0.126	0.355
DGATE	0.79	0.103	0.022	0.149	0.753	0.07	0.009	0.094	0.917	0.265	0.094	0.306
FONET	0.969	0.063	0.007	0.082	0.919	0.078	0.013	0.112	0.816	0.196	0.085	0.292
INDES	0.654	0.072	0.008	0.088	0.84	0.059	0.005	0.073	0.189	0.023	0.078	0.28
INGRM	0.745	0.115	0.022	0.147	0.662	0.117	0.035	0.186	0.54	0.316	0.118	0.343
KAREL	0.869	0.11	0.015	0.123	0.924	0.091	0.013	0.116	0.814	0.303	0.157	0.397
KRONT	0.723	0.103	0.015	0.124	0.857	0.093	0.019	0.136	0.487	0.218	0.077	0.278
LINK	0.973	0.061	0.007	0.086	0.937	0.063	0.007	0.085	0.983	0.336	0.24	0.49
LOGO	0.961	0.081	0.009	0.095	0.886	0.094	0.014	0.118	0.657	0.365	0.165	0.407
NETAS	0.913	0.019	0.001	0.03	0.773	0.044	0.003	0.056	0.484	0.2	0.096	0.311
PKART	0.806	0.082	0.019	0.137	0.947	0.09	0.012	0.111	0.725	0.241	0.13	0.361

The numerical results are presented in Figure 2 (R^2 values).

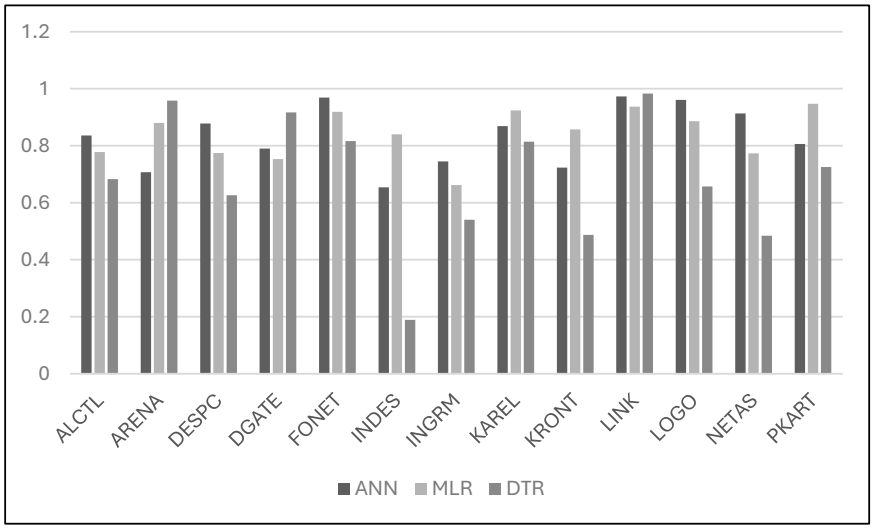


Figure 2. R^2 Values of models for ROA.

In general, the ANN model provides a good fit with high R^2 values for most companies. The highest R^2 value of ANN is 0.973 for LINK, and the lowest R^2 value is 0.654 for INDES. The MLR model also performs quite well and generally has high R^2 values. The highest R^2 value of MLR is 0.947 for PKART, and the lowest R^2 value is 0.662 for INGRM. The DTR model shows a wider range of R^2 values and more variability in its performance. The highest R^2 value of DTR is 0.983 for LINK, while the lowest R^2 value is 0.189 for INDES. In conclusion, the ANN model performs consistently well for most companies, with good generalization and predictive power. The MLR model also performs strongly, outperforming ANN in some cases (e.g., KAREL and PKART). The DTR model shows the highest R^2 values for some companies (e.g., LINK) and the lowest performance for others (e.g., INDES). This shows that the DTR model is more sensitive to the data range.

Figure 3 shows the RMSE values of companies' ROA values according to three different machine learning methods.

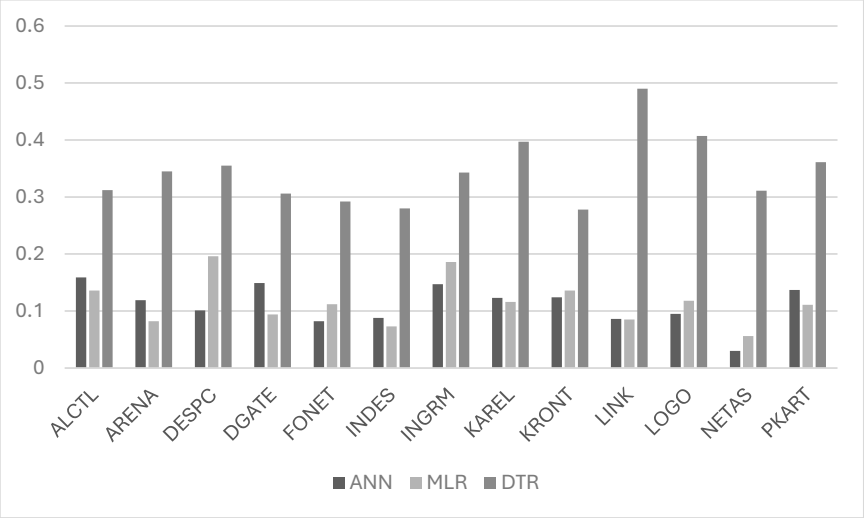


Figure 3. RMSE values of models for ROA.

In ANN, MLR and DTR, RMSE is the square root of the mean squared difference between the predicted values of a model and the actual values. In this context, lower RMSE values indicate better performance. The ANN model shows low RMSE values in general and exhibits a good prediction performance. The lowest RMSE value of the ANN is for NETAS (0.03), and the highest RMSE value is for ALCTL (0.159). The MLR model also performs well with low RMSE values in general. The lowest RMSE value of MLR is for ARENA (0.082), and the highest RMSE value is for DESPC (0.196). The DTR model, on the other hand, generally shows higher RMSE values, indicating that the model's forecasting performance is more variable. The lowest RMSE value of DTR is for FONET (0.292), and the highest RMSE value is for LINK (0.49). In general, the ANN and MLR models provide more consistent and reliable predictions with lower RMSE values, while the DTR model exhibits a more variable performance, showing higher RMSE values in some cases. These results suggest that ANN and MLR models generally offer better prediction performance, while the DTR model is more sensitive to the data structure.

The results of the models using machine learning methods for ROE are presented in Table 5.

Table 5. Estimation results for ROE.

	ANN				MLR				DTR			
	R^2	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE
ALCTL	0.847	0.292	0.143	0.179	0.503	0.178	0.045	0.212	0.913	0.17	0.05	0.224
ARENA	0.945	0.086	0.015	0.123	0.89	0.084	0.01	0.099	0.654	0.298	0.152	0.39
DESPC	0.95	0.094	0.014	0.118	0.908	0.108	0.018	0.133	0.998	0.269	0.132	0.363
DGATE	0.714	0.155	0.037	0.193	0.967	0.045	0.003	0.255	0.906	0.317	0.163	0.404
FONET	0.875	0.097	0.012	0.108	0.879	0.096	0.012	0.107	0.716	0.28	0.145	0.381
INDES	0.674	0.151	0.038	0.196	0.561	0.154	0.049	0.22	0.753	0.286	0.149	0.386
INGRM	0.437	0.177	0.043	0.207	0.837	0.273	0.113	0.336	0.916	0.291	0.118	0.343
KAREL	0.938	0.058	0.005	0.071	0.376	0.122	0.021	0.144	0.537	0.267	0.09	0.301
KRONT	0.925	0.059	0.005	0.07	0.651	0.112	0.027	0.164	0.991	0.25	0.126	0.356
LINK	0.912	0.096	0.022	0.127	0.947	0.112	0.017	0.132	0.897	0.218	0.104	0.323
LOGO	0.995	0.076	0.008	0.09	0.993	0.084	0.008	0.091	0.797	0.206	0.058	0.241
NETAS	0.411	0.109	0.045	0.153	0.252	0.131	0.038	0.196	0.451	0.111	0.047	0.216
PKART	0.984	0.051	0.004	0.065	0.977	0.049	0.005	0.067	0.613	0.156	0.056	0.236

The numerical results are presented in Figure 4 (R^2 values).

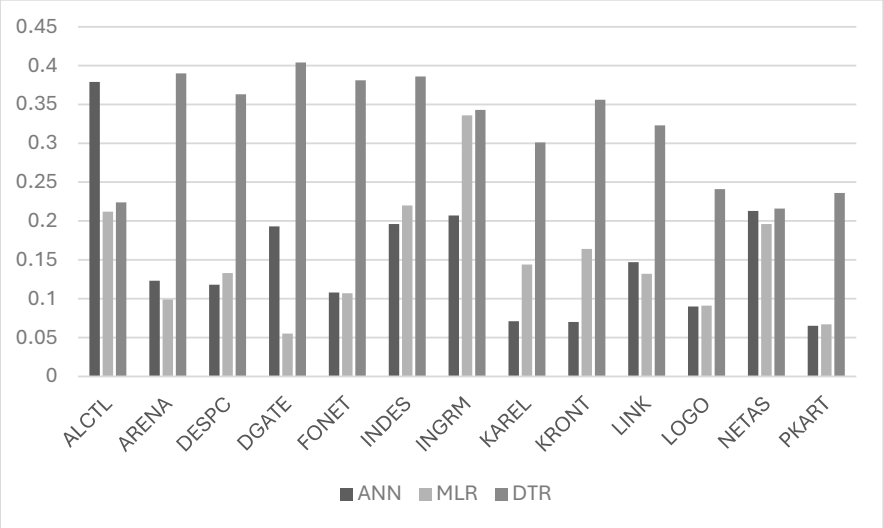


Figure 4. R^2 Values of models for ROE.

The R^2 values of the ANN model are generally high, indicating that it explains the variance of ROE better. For example, the R^2 value of the ANN for LOGO shows the highest value with 0.995. On the other hand, the R^2 value of the ANN for NETAS shows the lowest performance with 0.411. The multiple linear regression (MLR) model also performs well with generally high R^2 values. It has particularly high R^2 values for DGATE (0.967) and PKART (0.977). However, the MLR model performs more poorly, with low R^2 values for KAREL (0.376) and NETAS (0.252). The decision tree regression (DTR) model shows quite high R^2 values in some cases (e.g., 0.998 for DESPC and 0.991 for KRONT), while in other cases, it performs poorly (e.g., 0.654 for ARENA and 0.451 for NETAS). In general, ANN and MLR models generally explain the variance of the dependent variable well and provide reliable predictions with high R^2 values. The DTR model, on the other hand, may exhibit high performance in some cases, but it is more sensitive to the data structure and shows more variability in its performance. These results suggest that ANN and MLR models are

more consistent and reliable, while the DTR model may outperform them in certain cases. DTR is effective in capturing non-linear relationships in the data set. If the relationship between the dependent variable and independent variables is non-linear, DTR models can better capture this complexity and perform better. For example, the high R^2 values of DTR models for DESPC and KRONT (0.998 and 0.991) may indicate the presence of such non-linear relationships.

Figure 5 shows the RMSE values of companies' ROE values according to three different machine learning methods.

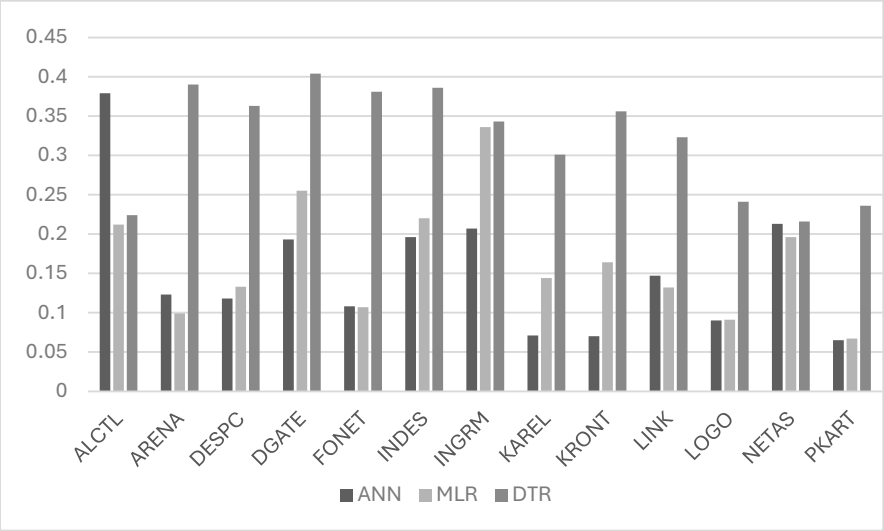


Figure 5. RMSE values of models for ROE.

The results obtained reveal that the overall performance of the methods is effective. When the ROE RMSE results are analyzed, the MLR method performed the best by reaching the lowest RMSE values in most firms. For example, in ARENA (0.099), FONET (0.107) and LOGO (0.091), the MLR method had significantly lower error rates than other methods. This superior performance of the MLR method shows that it works effectively especially in data sets where linear relationships are dominant. On the other hand, the ANN method outperforms MLR in some firms by providing lower RMSE values compared to MLR. For example, the ANN method achieved the lowest error rates in KAREL (0.071) and KRONT (0.07). The success of ANN in these firms reveals that it is more effective in data sets with complex and non-linear relationships. However, in some cases, the ANN method may have higher error rates, especially due to the complexity of the data sets. The DTR method, on the other hand, generally yielded the highest RMSE values and underperformed the ANN and MLR methods in most firms. The DTR method caused high error rates, especially in cases of noise and extreme complexity in the data sets. This may be due to the fact that the decision trees of the DTR method are prone to overfitting. In conclusion, the MLR method generally performed the best with the lowest RMSE values and stands out as the most appropriate forecasting method for many firms. The ANN method, on the other hand, outperformed others for some firms and is particularly effective in data sets with nonlinear relationships. The DTR method, on the other hand, showed the lowest performance with higher error rates in general. These results reveal that the performance of forecasting methods varies considerably depending on firm characteristics and the structure of the data sets.

5. Conclusions

The IT sector is a horizontal sector that affects all industries in an economy and plays a facilitating and productivity enhancing role. In a world moving from an industrial society where capital is important to an information society where information is important,

productivity and innovation are the key drivers of growth. IT plays a role in paving the way for the development of existing and emerging technologies and business areas through innovations, and increasing productivity and competitiveness. Within the scope of this definition, the sector has recently become one of the most active sectors for investors and policy makers. This interest stems from the fact that the sector is at the heart of innovation, competitiveness and economic growth. Given the rapidly evolving dynamics of the IT sector and the profitability-oriented strategies of investors, it is clear that accurate and reliable forecasts are critical for managers.

This study aims to predict ROA and ROE with machine learning methods using the financial data of 13 technology firms listed in the Borsa Istanbul Technology index and operating between March 2000 and December 2023. The results obtained using artificial neural networks, multiple linear regression and decision tree regression methods show that these methods are effective on profitability forecasting. The ROA forecasting results reveal that ANN and MLR methods perform better, especially ANN with overall higher R^2 and lower error metrics (MAE, MSE, RMSE). These findings emphasize the effectiveness of machine learning methods in predicting the financial performance of companies. The obtained ROE prediction results show that ANN and DTR methods perform better overall, with higher R^2 and lower error metrics (MAE, MSE, RMSE). These findings emphasize the effectiveness of machine learning methods in predicting the financial performance of companies.

In conclusion, profitability forecasting with machine learning methods in the IT sector has revealed important findings from both academic and practical perspectives and opened new doors for future research. The use of such methodologies in analyzing the financial performance of technology firms can be considered an important step in understanding the changes and innovations in the sector. This study can be a valuable resource for financial analysts and investors, contributing to more informed future investment decisions in the sector. In future studies, in-depth research can be conducted in the area of profitability forecasting in the IT sector. In particular, the following topics can be focused on. Comparison of Different Machine Learning Methods: Different machine learning methods such as artificial neural networks, decision tree regression and multiple linear regression can be compared and their effectiveness on profitability forecasting. It is important to determine which method is more successful in which situations. Analyzing Different Financial Indicators: In addition to ROA and ROE, the impact of other financial indicators (e.g., net profit margin, return on equity) on profitability forecasting can be examined. How the use of multiple indicators affects forecasting accuracy can be investigated. Expanding Existing Data Sets: Evaluating different time periods or a larger sample of firms beyond the data set used in the study may increase the generalizability of forecasting models. Comparison of Different Sectors: Comparative studies can be conducted on forecasting the financial performance of firms in different industries other than the IT sector. In this way, the impact of sectoral differences and special conditions on forecasting models can be understood.

Author Contributions: Conceptualization, B.T., M.K., M.T., H.T. and F.Y.; Methodology, M.K., M.T., G.F.Ü.U. and F.Y.; Formal analysis, M.K.; Resources, B.T. and G.F.Ü.U.; Data curation, H.T.; Writing—original draft, B.T., M.K., M.T. and G.F.Ü.U.; Writing—review & editing, H.T. and F.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Penman, S.H. *Financial Statement Analysis and Security Valuation*, 5th ed.; McGraw-Hill Education: New York, NY, USA, 2013.
2. Damodaran, A. *Corporate Finance: Theory and Practice*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2007.
3. Brigham, E.F.; Ehrhardt, M.C. *Financial Management: Theory & Practice*, 15th ed.; Cengage Learning: Boston, MA, USA, 2017.
4. Higgins, R.C. *Analysis for Financial Management*, 10th ed.; McGraw-Hill Education: New York, NY, USA, 2012.
5. Ross, S.A.; Westerfield, R.W.; Jaffe, J. *Corporate Finance*, 11th ed.; McGraw-Hill Education: New York, NY, USA, 2016.
6. Pham, V.H.S.; Le, T.D. Research on Applying Machine Learning Models to Predict and Assess Return on Assets (Roa). *Asian J. Civ. Eng.* **2024**. [CrossRef]
7. De Lucia, C.; Pazienza, P.; Bartlett, M. Does Good Esg Lead to Better Financial Performances by Firms? Machine Learning and Logistic Regression Models of Public Enterprises in Europe. *Sustainability* **2020**, *12*, 5317. [CrossRef]
8. Chen, T.-H.; Chang, R.-C. Using machine learning to evaluate the influence of FinTech patents: The case of Taiwan's financial industry. *J. Comput. Appl. Math.* **2021**, *390*, 113215. [CrossRef]
9. Rai, P.; Mohapatra, B.B.; Meitei, A.J.; Jain, V. Major Determinants of Bank Profitability in India: A Machine Learning Approach. *Glob. Bus. Rev.* **2023**, 09721509231184763. [CrossRef]
10. Chakri, P.; Pratap, S.; Gouda, S.K. An Exploratory Data Analysis Approach for Analyzing Financial Accounting Data Using Machine Learning. *Decis. Anal. J.* **2023**, *7*, 100212. [CrossRef]
11. Jones, S.; Moser, W.J.; Wieland, M.M. Machine Learning and The Prediction of Changes in Profitability. *Contemp. Account. Res.* **2023**, *40*, 2643–2672. [CrossRef]
12. Belesis, N.D.; Papanastasopoulos, G.A.; Vasilatos, A.M. Predicting the Profitability of Directional Changes Using Machine Learning: Evidence from European Countries. *J. Risk Financ. Manag.* **2023**, *16*, 520. [CrossRef]
13. Dong, X.; Dang, B.; Zang, H.; Li, S.; Ma, D. The Prediction Trend of Enterprise Financial Risk Based on Machine Learning Arima Model. *J. Theory Pract. Eng. Sci.* **2024**, *4*, 65–71.
14. Zhang, C.; Zhang, H.; Liu, D. A Contrastive Study of Machine Learning on Energy Firm Value Prediction. *IEEE Access* **2019**, *8*, 11635–11643. [CrossRef]
15. Zahariev, A.; Angelov, P.; Zarkova, S. Estimation of Bank Profitability Using Vector Error Correction Model and Support Vector Regression. *Econ. Altern.* **2022**, *28*, 157–170.
16. Kristof, T.; Virag, M. What Drives Financial Competitiveness of Industrial Sectors in Visegrad Four Countries? Evidence by Use of Machine Learning Techniques. *J. Compet.* **2022**, *14*, 117–136.
17. Anand, V.; Brunner, R.; Ikegwu, K.; Sougiannis, T. Predicting Profitability Using Machine Learning. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3466478 (accessed on 22 August 2024).
18. Nur-Al-Ahad, M.; Syeda, N.; Vagavi, P. Nexus Between Corporate Governance and Firm Performance in Malaysia: Supervised Machine Learning Approach. *Financ. Mark. Inst. Risks* **2019**, *3*, 115–130. [CrossRef]
19. Kayakuş, M.; Terzioglu, M.; Yetiz, F. Forecasting housing prices in Turkey by machine learning methods. *Aestimum* **2022**, *80*, 33–44. [CrossRef]
20. Agatonovic-Kustrin, S.; Beresford, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharm. Biomed. Anal.* **2000**, *22*, 717–727. [CrossRef]
21. Ahamed, K.I.; Akthar, S. A study on neural network architectures. *Comput. Eng. Intell. Syst.* **2016**, *7*, 1–7.
22. Iqbal, A.; Aftab, S. A feed-forward and pattern recognition ANN model for network intrusion detection. *Int. J. Comput. Netw. Inf. Secur.* **2019**, *14*, 19. [CrossRef]
23. Wang, W.Q.; Li, M.J.; Guo, J.Q.; Tao, W.Q. A feedforward-feedback control strategy based on artificial neural network for solar receivers. *Appl. Therm. Eng.* **2023**, *224*, 120069. [CrossRef]
24. Hoeting, J.; Raftery, A.E.; Madigan, D. A method for simultaneous variable selection and outlier identification in linear regression. *Comput. Stat. Data Anal.* **1996**, *22*, 251–270. [CrossRef]
25. Uyanık, G.K.; Güler, N. A study on multiple linear regression analysis. *Procedia-Soc. Behav. Sci.* **2013**, *106*, 234–240. [CrossRef]
26. Kantar, Y.M. Generalized least squares and weighted least squares estimation methods for distributional parameters. *REVSTAT-Stat. J.* **2015**, *13*, 263–282.
27. Ratner, B. Variable selection methods in regression: Ignorable problem, outing notable solution. *J. Target. Meas. Anal. Mark.* **2010**, *18*, 65–75. [CrossRef]
28. Nimon, K.F.; Oswald, F.L. Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organ. Res. Methods* **2013**, *16*, 650–674. [CrossRef]
29. Naidu, M.S.; Geethanjali, N. Classification of defects in software using decision tree algorithm. *Int. J. Eng. Sci. Technol.* **2013**, *5*, 1332.
30. Supsermpol, P.; Huynh, V.N.; Thajchayapong, S.; Chiadamrong, N. Geçiş döneminde Tayland'daki halka açık şirketlerin finansal performansının tahmini: Lojistik regresyon ve rastgele orman algoritması kullanan sınıf tabanlı bir yaklaşım. *Açık İnovasyon Derg. Teknol. Pazar Ve Karmaşıklık* **2023**, *9*, 100130.
31. Malerba, D.; Esposito, F.; Ceci, M.; Appice, A. Top-down induction of model trees with regression and splitting nodes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 612–625. [CrossRef]
32. Song, Y.Y.; Ying, L.U. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130.
33. Zhao, Y.; Zhang, Y. Comparison of decision tree methods for finding active objects. *Adv. Space Res.* **2008**, *41*, 1955–1959. [CrossRef]

34. Gajowniczek, K.; Ząbkowski, T. ImbTreeEntropy and ImbTreeAUC: Novel R packages for decision tree learning on the imbalanced datasets. *Electronics* **2021**, *10*, 657. [CrossRef]
35. Kayakuş, M.; Açıkgöz, F.Y. Classification of news texts by categories using machine learning methods. *Alphanumeric J.* **2022**, *10*, 155–166. [CrossRef]
36. Yağmur, A.; Kayakuş, M.; Terzioglu, M. House price prediction modeling using machine learning techniques: A comparative study. *Aestimum* **2022**, *81*, 39–51. [CrossRef]
37. Park, J.; Lee, W.H.; Kim, K.T.; Park, C.Y.; Lee, S.; Heo, T.Y. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. *Sci. Total Environ.* **2022**, *832*, 155070. [CrossRef] [PubMed]
38. Carlson, R.E.; Foley, T.A. The parameter R2 in multiquadric interpolation. *Comput. Math. Appl.* **1991**, *21*, 29–42. [CrossRef]
39. Hodson, T.O. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev. Discuss.* **2022**, *2022*, 1–10. [CrossRef]
40. Wei, Q.; Dunbrack, R.L., Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS ONE* **2013**, *8*, e67863. [CrossRef]
41. Dagum, C. The generation and distribution of income, the Lorenz curve and the Gini ratio. *Econ. Appliquée* **1980**, *33*, 327–367. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.