**Estimation of Independent Audit Company Preference with Machine Learning Methods**

**Abstract**

*This study aimed to estimate independent audit firm preferences of the enterprises in the Borsa Istanbul Yildiz Market in Turkey by using financial ratios and machine learning algorithms. In this context, 13 financial datasets of 159 enterprises traded in Borsa Istanbul in the 2019-2021 period were used as input variables. First, the significance values of the input variables were found by using the Mutual Information (MI) method. Then, input variables were grouped sequentially in order of importance to select the most accurate subset representing the data. Among the machine learning algorithms, SVM (Support Vector Machine), DT (Decision Tree), RF (Random Forest), NB (Naive Bayes), KNN (K-Nearest Neighbors), and XGBoost algorithm methods were used for group selection. GridSearchCV (GridSearch +Cross-Validation) technique was applied to optimize the initial parameters of the methods. As a result of the experiments, the XGBoost algorithm was found to be the most successful method in the preference of independent audit companies with an accuracy value of 88.4%. It was sufficient for the method to use 8 attributes selected from 13 datasets. On the other hand, the Assets Profitability dataset was determined as the most important attribute. The study both used the MI method in attribute selection and estimated the preference for an independent audit firm with high success by removing less important variables from the model*.

***Keywords:*** *Independent Audit, Financial Ratios, Machine Learning, Classification*

**1. Introduction**

In Turkey, there is no obligation for public companies to inform the public about the factors that they are influenced by in their independent audit company preferences other than mandatory rotation.  This constitutes a major deficiency for information users such as creditors, investors, and brokerage houses that want to have an idea about issues such as company performance, audit quality, etc. Although there are many factors affecting audit firm preference in the literature, it has been observed that studies have mostly focused on financial performance factors. Table 1 lists the studies explaining the audit company change with financial performance and the methods they used.

**Table 1.** Summary of studies on the relationship between audit firm change and financial performance

|  |  |  |
| --- | --- | --- |
| **Author(s)** | **Methods used** | **Result(s)** |
| Dhaliwal (2007) | Logistic regression | It was argued that the management of a company that displays poor financial performance would prefer an audit company that would make the financial status of the company look better. |
| Chadegani et al. (2011) | Logistic regression | They found that one of the 6 factors affecting the independent audit company change was financial failure. |
| Kwak et al. (2011) | Bayesian network | It was determined that in companies that change the audit company, the variables of Net Working Capital / Total Assets, Short-Term Debts / Current Assets, Dividend Payment, Profit Before Interest and Tax / Total Assets were significantly different compared to companies that do not change the audit companies. |
| Nazri et al. (2012) | Survey, Logistic regression | The variables included in the study consisted of elements such as change in management, negative audit reports, business size, audit quality, financial structure, and audit fee. As a result of the study, it was found that the increase in the level of indebtedness increased the likelihood of the client business to change the audit firm. |
| Huang and Scholz (2012) | Panel data regression | They found that businesses that counterfeit financial data (turning profits into losses, etc.) caused their audit firms to resign and then they tried to negotiate with larger audit firms. |
| Eldridge et al. (2012) | Logistic discriminant | It was attempted to estimate audit firm change based on 13 financial failure variables. The variables that were effective in estimating the changes of the audit firm were found to be profit share, loss, undistributed profit/total asset, and profit before interest and tax/ total asset. |
| Suyono et al. (2013) | Questionnaire survey | A significant and positive relationship was determined between the financial performance of the business and audit firm change. The most important indicators were found as financial performance, competition between audit firms, and audit quality. |
| Black et al. (2013) | Logistic regression, logistic discriminant | It was determined that 4 financial ratios (Net Working Capital/Total Assets, Return on Assets, Equity/Total Debt, Sales/Total Assets) were effective in explaining the audit company change. In addition, it was found that while the success of the logistic regression method in accurately classifying the audit company change was 92.3%, the success of the discriminant analysis was 47.6%. |
| Aroh et al. (2017) | Logistic regression | No positive relationship was found between the financial failure variable and the audit firm change. |
| Kusuma and Farida (2019) | Discriminant, logistic regression | They found that financial failure, profitability, and management change had a significant impact on audit firm change. |
| Dharmasari and Suardana (2021) | Logistic regression | They analyzed the financial statements of 25 manufacturing enterprises registered in the Indonesian stock exchange and found a positive relationship between financial weakness and audit firm change. |

In this context, it is thought that it would be effective to estimate the audit firm preference of public companies in Turkey with the help of financial ratios. When the studies conducted in the literature were evaluated in general, it was determined that statistical methods such as questionnaires, question papers, logistic regression, and discriminant analysis were mostly used. Since no studies explaining this relationship by using machine learning algorithms were encountered, it was thought that estimating the relationship between independent audit company preference and financial ratios with machine learning algorithms such as SVM (Support Vector Machine), DT (Decision Tree), RF (Random Forest), NB (Naive Bayes), KNN (K-Nearest Neighbors), and XGBoost would fill the gap in the literature in this regard.

**2.** **Methodology**

The learning process includes processes such as classification, regression, clustering, and attribute extraction. Learning from data is defined as the creation of a learning machine or algorithm that allows the estimation function to be determined based on a limited number of training data. Owing to these characteristics, learning with data-based methods (data mining) is used in many fields such as marketing, finance, management, engineering, production, industry, banking, health, tourism, and medicine (Chunhong and Licheng, 2004, p.1869). In the process of separating these data sets into certain classes according to their common characteristics, various classification methods are needed. Various algorithms have been developed for this purpose. The main algorithms used in this regard can be listed as entropy-based classification (C4.5 algorithm, C5.0 algorithm), Regression and Decision Trees (Gini algorithm, Twoing algorithm), Memory-Based Algorithms (K-nearest neighbor algorithm), Bayesian Classifiers, Regression Trees, Support Vector Machines, and Random Forest (Oleg, 2010, p.11).

Among these algorithms, to determine whether an effective estimation of the audit firm preference would be made and to identify the significance levels of the determined data sets (financial ratios), the relationship between the dependent variable (audit firm preference) and the input variables (financial ratios) has been explained using Support Vector Machines, Decision Trees, Random Forest, Naive Bayes, K-Nearest Neighbor, and XGBoost classification algorithms. The purpose here is to establish an estimation model that predicts the relationship between dependent and input variables to have the best fit. Descriptive information about the machine learning methods used in the research are presented below.

**2.1. Support Vector Machines (SVM):**When they were developed in the 1990s, they became extremely popular and became one of the best methods for a high-performance algorithm (Pradhan, 2012, p.82). Support vector machines are a supervised machine learning algorithm with strong foundations based on the Vapnik-Chervonenkis theory. SVM is similar to neural networks and radial-based artificial neural networks, but it is generally indicated that it performs better than these algorithms. This learning algorithm is seen as an alternative training technique for Polynomial, Radial Basic Function and Multilayer Perceptron (Artificial neural network) classifiers. The underlying idea in the technique is to create an optimal hyperplane that maximizes the margin between classes of data (Pradhan, 2012, p.83). SVM is particularly useful in terms of its simple structure and high performance in practical applications. The number of samples to be used in SVMs is not important. SVM also has the ability to generalize data that has not been seen during training, as it classifies them as problem-free. Its ability to generalize makes SVM a good alternative to other techniques. SVMs can be run on classification and regression problems. Therefore,it has been observed that they are more successful in comparison to statistical methods in various areas such as predicting stock market index movements, predicting financial information manipulations, or creating early warning systems against financial failures(Cao and Tay, 2003, p.1506-1507).

The main idea in the SVM method is that there is a linear discriminating function that reflects the character of the available training data as close to reality as possible and fits the statistical learning theory. Since non-linear situations can be processed, the mapping function that provides dimension change is not known, and it is difficult to process in high dimensions, arrangements called kernel tricks are made (Hoffman, 2006, p.4). Thus, instead of the mapping function in the transformed space, the kernel functions that ensure the use of data directly in the input space are included in the process. There are four kernel functions commonly used in the Support Vector Machines (SVM) algorithm. These are (Nanda et al., 2018, p.6):

* Linear Function,
* Radial Basis Function,
* Polynomial Function, and
* Sigmoid Function.

Instead of calculating the product values of all values repeatedly by using kernel functions, it is ensured that the value in the quality space is found by replacing the value directly in the kernel function. In this way, there is no possibility of dealing with an extremely high-dimensional quality space.

**2.2. Decision Tree (DT):** It is a data mining approach that is generally used for classification and estimation (Song and Lu, 2015, p.130). Although it is used for classification in other methods such as artificial neural network, it has many advantages such as being easy to interpret and understand by comparing its own decisions with its domain information in order to make verification (Nassif et al., 2013, p.220). Moreover, it allows the analysis of various data without requiring assumptions. The first step to take in creating a decision tree is to determine according to which criterion the branching in the tree will be made or according to which attribute values the tree structure will be created (Seethapathy and Babu, 2021, p.13). For this purpose, a variety of algorithms have been developed in the literature. C4.5, C5.0, CHAID, CART, and ID3 can be given as examples of these (Patel and Prajapati, 2018, p.75; Seethapathy and Babu, 2021, p.13-14). Decision trees are commonly used as they can be easily implemented in integration with databases and are reliable (Nassif et al., 2013, p.220). The method is based on investigating all possible relationships between dependent and independent variables to obtain the strongest and best estimate (Seethapathy and Babu, 2021, pp.14-15). When the independent variable with the strongest relationship is detected, the dataset is divided into two according to the values of this independent variable. This process is continued until the possible divisions are completed (Gupta et al., 2017, p.16). The decision rules reflected by the logical model presented by decision trees are clear enough to be easily understood by humans. This method has a wide range of applications since it possesses features such as high classification accuracy rate and simple rules produced (Barros et al., 2012, p.1237).

**2.3. Random Forest (RF) algorithm:** It is a method consisting of decision trees and nodes developed by Breiman (2001). With this method, regression and classification analyses can be performed (Liaw and Wiener, 2022, p.18). In this method, the best random values in the nodes are selected to divide the nodes into branches, and certain weights are assigned to the decision trees created (Siva et al., 2012, p.133). These weights are determined in line with the internal errors of the decision trees, and the decision tree with the lowest error is assigned the highest weight, and the decision tree with the highest error is given the lowest weight. These weights assigned are used for voting in class estimation. Then, these votes are collected, and the ultimate decision is made (Chen et al., 2009, p.244).

**2.4. Naive Bayes algorithm (Naive Bayes-NB):** The classifier algorithm, named as the Gausssan Naive Bayes classifier, or commonly referred to simply as Naive Bayes, is a simple probability-based classification method based on the Bayes theorem. The Naive Bayes method is a simple but powerful algorithm for creating a prediction-based modeling. Therefore, it is one of the most used classification and prediction algorithms, especially in the fields of signal and image processing. In this algorithm, it is assumed that the presence of certain features in a class has no relationship with any other feature (Srinivasa et al., 2020, p.7594).

**2.5. XGBoost algorithm (XGBoost):** The XGBoost algorithm was developed by Tianqi Chen and Carlos Guestrin as a research project at the University of Washington in 2016 (Chen et al., 2016, p.789). The XGBoost algorithm is a decision tree-based, ensemble learning-based machine learning algorithm that uses the gradient enhancement framework. The algorithm has attracted attention as it has won many Kaggle competitions. XGboost is used in structured or tabular data sets in regression and classification problems (Zhou et al., 2021, p.28). The XGBoost algorithm uses the CART (Classification and Regression Tree) algorithm to form trees. XGBoost is an improved implementation of the basic Gradient Boosting Machines framework through system optimization and algorithmic improvements (Chen et al., 2016, p.789).

**2.6. K-Nearest Neighbors algorithm:** The main purpose of classification is to determine which class the objects belong to by looking at the properties of the objects. Many different classification types and algorithms are available. Decision trees, nearest neighbors, Bayes, and artificial neural networks are some of them. The KNN algorithm, also known as the K-Nearest Neighbor algorithm, is one of the most well-known and used algorithms among machine learning algorithms (Hu et al., 2016, p.7). Classification is made by using the closeness between the features closest to a selected feature. The K value found here is expressed by a number, for example, 3 or 5. When we look at the way it functions, the K value is first checked when an object that needs to be defined according to the defined data emerges. In order to avoid equality here, the number K is usually selected as an odd number. When calculating the distances between the new data and other data, methods such as Cosine, Euclidean, or Manhattan distance are used (Han et al., 2011, p.45).

**3. Dataset**

In this environment, where the full application of the principle of financial transparency, which is of great importance for information users such as investors, lenders, and regulatory bodies, in the markets is discussed, there is no information sharing about the factors according to which the independent audit company preferences of the enterprises are made. Due to this gap, a literature review was conducted on which factors businesses are affected in their independent audit company preferences, and although many factors were identified, it was observed that the weighted point of the studies was audit fee and financial performance factors ( Nasser et al., 2006; Wan Mohamed et al., 2007; Ettredge et al., 2007; Ismail et al., 2008; Calderon and Ofobike, 2008; Chen et al., 2008; Kwak et al., 2011; Eldridge et al., 2012; Huang and Scholz, 2012; Suyono et al., 2013; Black et al., 2013). However, as the data on audit fees in Turkey only show the revenue generated by the audit firms from the total audit of that year and the audit fee is not reflected on the client business basis, the present study focused on the financial performance variable. In this respect, the present study aimed to estimate the factors affecting the independent audit company preferences of the companies traded in the Turkish Stock Exchange Istanbul (BIST) Star Market by using Machine Learning algorithms. Information about the calculation of 1 dependent, 13 input variables created for use in machine learning algorithms is presented in Table 2.

**Table 2.** Variables Used in the Model

|  |  |  |
| --- | --- | --- |
| **Dependent Variable** | **Input Variables** | **Calculation** |
|  | Current Ratio (CR) | Current assets / K.V. financial payables |
|  | Liquidity Ratio (LR) | (Current assets-Inventories) / K. V. financial payables |
|  | Cash Ratio (CR) | (Current assets-Inventories-Receivables) /K.V. financial payables |
|  | Financial Debt Ratio (FDR) | (K.V. Financial Liabilities + U.V. Financial Liabilities) / Total Assets |
| Audit Firm Preference (AFP) | Leverage Ratio (LR) | Total Liabilities / Total Assets |
|  | Asset Turnover Rate (ATR) | Net sales /Total assets |
|  | Inventory Turnover Rate (ITR)  Return on Assets (RA)  Gross Profit Margin (GPM)  EBIDTA Margin (EM)  Net Profit Margin (NPM)  Return on Equity (ROI)  Return on Capital Investments (ROIC) | Cost of sales/ Average stock amount  Net Profit /Total Assets  Gross sales profit / Net sales  EBITDA / Total Sales  Net profit / Net sales  Net profit / Equity  Net Operating Profit After Tax / Capital Deposited |

**Source:** Saalem and Rehman, 2011, p.97; Adjirackor et al., 2017, p.6

While the dependent variable used in the study was defined as Audit Firm Preference (AFP), the input variables were determined as Asset Profitability (AP), Financial Debt Ratio (FDR), Net Profit Margin (NPM), Return on Investment (ROIC), Favök (EBITDA) Margin (FM), Leverage Ratio (LR), Stock Turnover Rate (STR), Current Ratio (CR), Gross Profit Margin (GPM), Cash Ratio (CR), Liquidity Ratio (LR), Asset Turnover Rate (ATR), and Equity Profitability (EP).

The data related to these input variables, which were thought to best represent the financial capabilities of the enterprises, were obtained from the website of Fintables Informatics Technologies Inc. which included the financial data analyses of the enterprises, covering the years 2019-2021. Since these input variables, which were determined in order to reveal the relationship between financial ratios and independent audit firm preference, were taken as indicators that could directly affect financial performance in many studies (e.g., Cheng et al., 2005; Lei & Liu, 2021; Yim & Mitcbell, 2005; Lin, 2009), it was deemed appropriate to use them in this study as well.

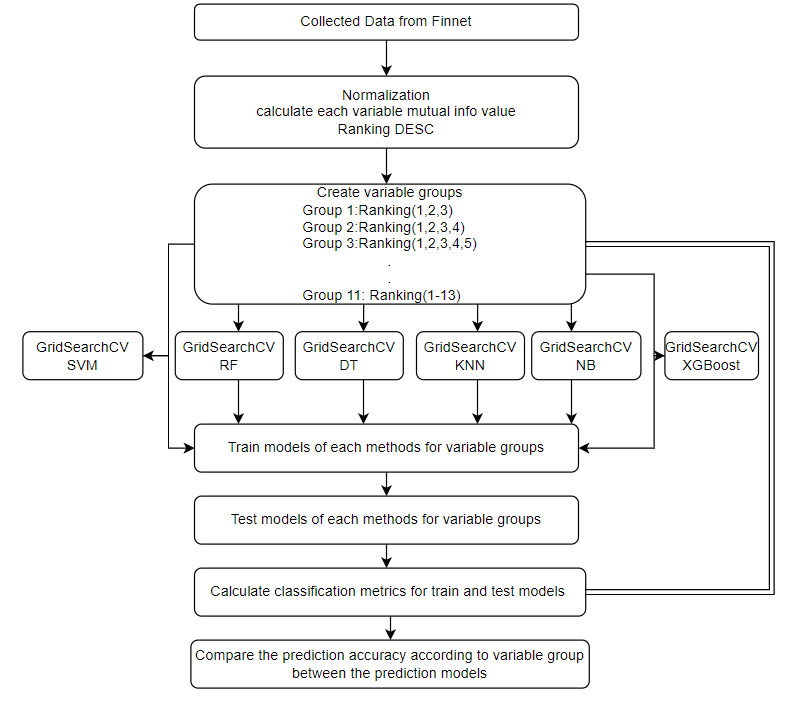
While classifying the dependent variable of the research, if the audit firm preference (AFP) of the said enterprises was one of the four major audit firms (Ernst and Young-Guney Independent Audit, PWC - Basaran Nas Independent Audit, Deloitte - DRT Independent Audit, KPMG-Akis Independent Audit) it was given the value of 1, while it was given the value of 0 for an audit firm other than the four major firms.

Sectors such as packaging, chemicals and plastic, metal goods and machinery, textile clothing and leather, transportation, manufacturing, base metal, defense, information and software, paper products, stone soil cement, furniture, and wholesale and retail trade, which are traded in BIST Yildiz Market, were included in the research. Banks, factoring companies, financial leasing companies, and investment trusts were excluded from the study due to the differences in their fields of activity and financial statement formats, as well as the fact that their asset structures were mostly cash and cash equivalents. Enterprises whose existence was terminated due to situations such as interrupted activities, bankruptcies, etc. between 2019-2021 were also removed from the data set.

Since classical statistical methods can be partially successful in the estimation phase, it was preferred to use SVM, RF, DT, KNN, NB, and XGBoost algorithms, which are among the machine learning methods that can learn from samples in recent years, can give a higher accuracy rate with fewer data sets, have generalization ability, and have fewer assumptions compared to statistical methods.

**4.** **Experiment and Results**

The study was coded using the Python 3.6 scikit-learn library. The data set of the research consisted of 477 data sets (159 companies x 3 years) between 2019-2021. 80% of the data belonging to the variable group were used as training data and 20% as test data. The process steps performed for the research and the information about these processes are presented in Figure 1.



**Figure 1.** The Process of the Experiment

As can be seen in Figure 1, the collected data were first normalized by the min-max normalization method to be used in classification. In this method, the smallest (0.01) and largest (0.99) values in a group of data are taken, and all other data are normalized in this value range (0.01-0.99) (Wu et al., 2016, p.219). This situation is mathematically formulated as (Arora et al., 2021, p.1337)

Input parameter value: Xi,t

New value: Xt

. Table 3 presents the descriptive statistics of the data set.

**Table 3.** Initial Data Set Descriptive Statistics

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CR** | **LR** | **CR** | **FDR** | **LR** | **ATR** | | **ITR** | **RA** | **GPM** | **EM** | **NPM** | **ROI** | **ROIC** |
| **Number** | 474 | 474 | 474 | 474 | 474 | | 474 | 474 | 474 | 474 | 474 | 474 | 474 | 474 |
| **Average** | 1.9 | 1.4 | 0.7 | 26.7 | 53.8 | | 0.9 | 9.9 | 8.9 | 25.5 | 16.0 | 11.2 | 19.8 | 22.5 |
| **Std.** | 1.9 | 1.8 | 1.6 | 20.4 | 25.2 | | 0.7 | 15.6 | 12.5 | 20 | 19.2 | 25.4 | 29.3 | 36.6 |
| **Min** | 0.1 | 0.09 | 0 | 0 | 0.9 | | 0 | -0.2 | -38.1 | -37.9 | -84.5 | -116 | -118 | -103.2 |
| **25%** | 1 | 0.6 | 0.09 | 9.03 | 35.8 | | 0.4 | 3.1 | 1.7 | 13.8 | 7.9 | 1.7 | 4.2 | 4.5 |
| **50%** | 1.3 | 0.9 | 0.2 | 25.5 | 57.0 | | 0.8 | 5.3 | 6.3 | 22.2 | 15.2 | 7.6 | 18.5 | 21 |
| **75%** | 2.1 | 1.5 | 0.6 | 39.6 | 72.7 | | 1.2 | 9.8 | 14.4 | 32.5 | 22.3 | 16.9 | 32.2 | 42 |
| **Max** | 14.6 | 14.5 | 17.3 | 158.8 | 260.5 | | 4.3 | 124.6 | 75 | 118.4 | 94.9 | 177.1 | 339.5 | 199 |

The mutual Information method was used to identify attributes with high importance levels in the data set. Mutual Information (MI) measures the amount of information that a random variable can obtain about the other random variable (Witten, 2016, p.310). Mathematically, mutual information is expressed for the two abstract random variables of x and y as (Peng et al., 2005, p.1226):

Here, p (x,y) shows relative density function,

for x and y, p(x) and p(y) represent marginal probability density function.

The purpose of using MI is that it is independent of any classifier and does not require any parameter setting. Therefore, it is easy to apply and has a good generalization ability for different data. The order of importance of the variables according to the attribute evaluation method MI is shown in Table 4. The variables consisted of 11 groups, with the first three variables according to the order of importance in Table 4 (RA, NPM, FDR) Ranking (1)-(3), (RA, NPM, FDR, ROIC) Ranking (1)-(4), … Ranking (1)-(13). Each group was assigned to the machine learning classification algorithms, namely SVM, RF, DT, KNN, NB, and XGBoost methods, as input variable groups to estimate the audit firm preference determined as the dependent variable.

**Table 4.** Mutual Information Importance Ranking of the Variables Used in the Study

|  |  |  |  |
| --- | --- | --- | --- |
| **Input Variable** | **Ranking** | **Mutual Information Value** | **Input Variable Group** |
| RA | 1 | 0.0580 |  |
| NPM | 2 | 0.0362 |  |
| FDR | 3 | 0.0300 | Ranking (1)-(3) |
| ROIC | 4 | 0.0284 | Ranking (1)-(4) |
| EM | 5 | 0.0256 | Ranking (1)-(5) |
| LR | 6 | 0.0211 | Ranking (1)-(6) |
| CR | 7 | 0.0210 | Ranking (1)-(7) |
| ITR | 8 | 0.0202 | Ranking (1)-(8) |
| ATR | 9 | 0.0063 | Ranking (1)-(9) |
| GPM | 10 | 0.0053 | Ranking (1)-(10) |
| ROI | 11 | 0.0023 | Ranking (1)-(11) |
| CR | 12 | 0.0000 | Ranking (1)-(12) |
| LR | 13 | 0.0000 | Ranking (1)-(13) |

|  |
| --- |
|  |
| **Figure 2.** Importance Value of Features |

Considering that there are many hyperparameters for machine learning algorithms and many values that these hyperparameters can take, it is necessary to choose the best parameter combination to be used in SVM, RF, DT, KNN, NB, and XGBoost methods. In this respect, especially when working with small data sets, GridSearchCV (GridSearch+Cross-Validation) hyperparameter optimization method, which gives successful results in determining the hyperparameter set with the best performance, was preferred (Shuai et al., 2018, s.451). In this method, separate models are set with all combinations for the hyperparameters and values desired to be tested, and the most successful hyperparameter set is determined according to the metrics established. The parameters and their potential values used in the GridSearchCV method in the study are presented in Table 5.

**Table 5.** Values Used in Parameter Optimization

|  |  |  |
| --- | --- | --- |
| **Models** | **Parameter** | **Parameter’s Value** |
| SVM | C  kernel  degree | 0.1,1,5,10,15,20,100,1000  rbf, linear, poly, sigmoid  3,8 |
| RF | n estimators  max depth  min samples leaf  min samples split  criterion  max features | 1,5,10,50,100  4,5,6,7,8,9,10  2,10,20  10,15,20  gini, entropy  auto, sqrt |
| DT | max leaf nodes  min samples leaf  min sample split  criterion  max depth  max features | 2-100  1-5  2,3,4  gini, entropy  2,3,4,5,6,7,8,9,10,11,12  auto, sqrt, log2 |
| KNN | n-neighbors  leaf\_size  p  weights  algorithm  metric | range (1,30)  20,40,1  1,2  'uniform', 'distance'  'auto', 'ball\_tree', 'kd\_tree', 'brute'  'Minkowski', 'chebyshev' |
| NB | Priors  var\_smoothing | [None, [0.1,]\* len (n\_classes),]  [1e-9, 1e-6, 1e-12] |
| XGBoost | Objective  learning\_rate  max\_depth  min\_child\_weight  subsample  colsample\_bytree  n\_estimators | ['binary: logistic']  [0.001, 0.01, 0.1, 0.20, 0.25, 0.30]  [3,4,5,6,8,10,12,15]  [1,5,10,11]  [0.8]  [0.7]  [5,100,500,1000] |

After the parameters required for the models were determined with the GridSearchCV method, with each of the SVM, RF, DT, KNN, NB, and XGBoost machine learning methods, 66 estimation models for 11 different variable groups were established. In or der to measure the success rate of the estimation models, the performance values used for classification such as accuracy, precision, recall, f1-score, and Cohen’s Kappa values were calculated and presented in Table 6.

**Table 6.** Prediction Result of Models According to the Variants of Input Variable

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | **Test** | | | | |
| **Model** | **Input Variable Group** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **Kappa** |
| SVM | Ranking (1)-(3) | 0.621 | 0.521 | 0.621 | 0.501 | -0.006 |
| Ranking (1)-(4) | 0.642 | 0.616 | 0.642 | 0.568 | 0.09 |
| Ranking (1)-(5) | 0.747 | 0.75 | 0.747 | 0.729 | 0.408 |
| Ranking (1)-(6) | 0.716 | 0.717 | 0.716 | 0.688 | 0.321 |
| Ranking (1)-(7) | 0.747 | 0.742 | 0.747 | 0.739 | 0.43 |
| Ranking (1)-(8) | 0.705 | 0.696 | 0.705 | 0.695 | 0.335 |
| Ranking (1)-(9) | 0.737 | 0.768 | 0.737 | 0.7 | 0.354 |
| Ranking (1)-(10) | 0.695 | 0.684 | 0.695 | 0.683 | 0.307 |
| Ranking (1)-(11) | 0.684 | 0.672 | 0.684 | 0.67 | 0.278 |
| Ranking (1)-(12) | 0.684 | 0.673 | 0.684 | 0.656 | 0.25 |
| Ranking (1)-(13) | 0.695 | 0.689 | 0.695 | 0.665 | 0.27 |
| RF | Ranking (1)-(3) | 0.653 | 0.632 | 0.653 | 0.625 | 0.18 |
| Ranking (1)-(4) | 0.726 | 0.735 | 0.726 | 0.697 | 0.341 |
| Ranking (1)-(5) | 0.716 | 0.711 | 0.716 | 0.713 | 0.378 |
| Ranking (1)-(6) | 0.758 | 0.753 | 0.758 | 0.751 | 0.457 |
| Ranking (1)-(7) | 0.758 | 0.757 | 0.758 | 0.745 | 0.443 |
| Ranking (1)-(8) | 0.768 | 0.765 | 0.768 | 0.761 | 0.478 |
| Ranking (1)-(9) | 0.789 | 0.791 | 0.789 | 0.78 | 0.519 |
| Ranking (1)-(10) | 0.789 | 0.787 | 0.789 | 0.786 | 0.537 |
| Ranking (1)-(11) | 0.705 | 0.698 | 0.705 | 0.698 | 0.343 |
| Ranking (1)-(12) | 0.811 | 0.808 | 0.811 | 0.808 | 0.583 |
| Ranking (1)-(13) | 0.779 | 0.776 | 0.779 | 0.777 | 0.516 |
| DT | Ranking (1)-(3) | 0.632 | 0.608 | 0.632 | 0.608 | 0.142 |
| Ranking (1)-(4) | 0.653 | 0.647 | 0.653 | 0.649 | 0.24 |
| Ranking (1)-(5) | 0.653 | 0.689 | 0.653 | 0.659 | 0.307 |
| Ranking (1)-(6) | 0.705 | 0.696 | 0.705 | 0.695 | 0.335 |
| Ranking (1)-(7) | 0.684 | 0.698 | 0.684 | 0.688 | 0.345 |
| Ranking (1)-(8) | 0.663 | 0.659 | 0.663 | 0.661 | 0.267 |
| Ranking (1)-(9) | 0.768 | 0.766 | 0.768 | 0.767 | 0.496 |
| Ranking (1)-(10) | 0.674 | 0.71 | 0.674 | 0.679 | 0.349 |
| Ranking (1)-(11) | 0.737 | 0.759 | 0.737 | 0.741 | 0.463 |
| Ranking (1)-(12) | 0.737 | 0.752 | 0.737 | 0.741 | 0.457 |
| Ranking (1)-(13) | 0.705 | 0.714 | 0.705 | 0.708 | 0.381 |
| NB | Ranking (1)-(3) | 0.642 | 0.614 | 0.642 | 0.588 | 0.115 |
| Ranking (1)-(4) | 0.663 | 0.647 | 0.663 | 0.62 | 0.178 |
| Ranking (1)-(5) | 0.684 | 0.673 | 0.684 | 0.656 | 0.25 |
| Ranking (1)-(6) | 0.737 | 0.733 | 0.737 | 0.734 | 0.424 |
| Ranking (1)-(7) | 0.579 | 0.619 | 0.579 | 0.586 | 0.165 |
| Ranking (1)-(8) | 0.579 | 0.671 | 0.579 | 0.578 | 0.216 |
| Ranking (1)-(9) | 0.568 | 0.653 | 0.568 | 0.568 | 0.193 |
| Ranking (1)-(10) | 0.579 | 0.66 | 0.579 | 0.58 | 0.208 |
| Ranking (1)-(11) | 0.568 | 0.664 | 0.568 | 0.566 | 0.201 |
| Ranking (1)-(12) | 0.558 | 0.657 | 0.558 | 0.554 | 0.186 |
| Ranking (1)-(13) | 0.547 | 0.65 | 0.547 | 0.542 | 0.171 |
| KNN | Ranking (1)-(3) | 0.663 | 0.653 | 0.663 | 0.655 | 0.249 |
| Ranking (1)-(4) | 0.716 | 0.708 | 0.716 | 0.708 | 0.363 |
| Ranking (1)-(5) | 0.726 | 0.721 | 0.726 | 0.722 | 0.398 |
| Ranking (1)-(6) | 0.747 | 0.743 | 0.747 | 0.744 | 0.444 |
| Ranking (1)-(7) | 0.726 | 0.721 | 0.726 | 0.722 | 0.398 |
| Ranking (1)-(8) | 0.653 | 0.664 | 0.653 | 0.657 | 0.275 |
| Ranking (1)-(9) | 0.789 | 0.788 | 0.789 | 0.782 | 0.525 |
| Ranking (1)-(10) | 0.811 | 0.811 | 0.811 | 0.804 | 0.572 |
| Ranking (1)-(11) | 0.779 | 0.775 | 0.779 | 0.775 | 0.51 |
| Ranking (1)-(12) | 0.779 | 0.776 | 0.779 | 0.773 | 0.504 |
| Ranking (1)-(13) | 0.768 | 0.764 | 0.768 | 0.763 | 0.484 |
| XGBoost | Ranking (1)-(3) | 0.684 | 0.684 | 0.684 | 0.684 | 0.321 |
| Ranking (1)-(4) | 0.716 | 0.748 | 0.716 | 0.669 | 0.292 |
| Ranking (1)-(5) | 0.737 | 0.735 | 0.737 | 0.72 | 0.387 |
| Ranking (1)-(6) | 0.768 | 0.771 | 0.768 | 0.755 | 0.464 |
| Ranking (1)-(7) | 0.758 | 0.753 | 0.758 | 0.751 | 0.457 |
| Ranking (1)-(8) | 0.884 | 0.884 | 0.884 | 0.884 | 0.75 |
| Ranking (1)-(9) | 0.874 | 0.874 | 0.874 | 0.872 | 0.722 |
| Ranking (1)-(10) | 0.853 | 0.853 | 0.853 | 0.853 | 0.683 |
| Ranking (1)-(11) | 0.821 | 0.819 | 0.821 | 0.819 | 0.608 |
| Ranking (1)-(12) | 0.863 | 0.863 | 0.863 | 0.863 | 0.704 |
| Ranking (1)-(13) | 0.811 | 0.808 | 0.811 | 0.808 | 0.583 |

The formulas used in the calculation of the performance metrics in Table 6 are presented below (Briliani, 2019, p.102-103).

The following formula was used to calculate the accuracy value.

*Accuracy=*

The following formula was used to calculate the precision value.

*Precision=*

The following formula was used to calculate the recall value.

*Recall=*

The following formula was used to calculate the F1-score value.

*F1-score=2\**

Cohen’s Kappa

Relative observed agreement among accuracy.

In the formulas, TP represents the real positive number, TN the real negative number, FP the incorrect positive number, and FN the incorrect negative number. In Figure 3, where the performance metrics of the machine learning algorithms according to the test results are presented, it was observed that the XGBoost algorithm yielded a more successful estimation result compared to the other machine learning methods.



**Figure 3.** Comparison of ML Algorithms According to Performance Metrics

Table 7 presents the comparison of the models according to the most successful variable groups.

**Table 7.** Comparison of the Models According to the Most Successful Variable Groups

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Cohen’s Kappa** | **Ranking Group** |
| XGBoost | 0.884 | 0.884 | 0.884 | 0.884 | 0.75 | Ranking (1)-(8) |
| RF | 0.811 | 0.808 | 0.811 | 0.808 | 0.583 | Ranking (1)-(12) |
| KNN | 0.811 | 0.811 | 0.811 | 0.804 | 0.572 | Ranking (1)-(10) |
| SVM | 0.747 | 0.75 | 0.747 | 0.729 | 0.408 | Ranking (1)-(5) |
| DT | 0.737 | 0.759 | 0.737 | 0.741 | 0.463 | Ranking (1)-(11) |
| NB | 0.737 | 0.733 | 0.737 | 0.734 | 0.424 | Ranking (1)-(6) |

When ranked according to performance metrics, the best performance was exhibited by the 8-variable (RA, NPM, FDR, ROIC, EM, LR, CR, ITR) XGBoost algorithm with an accuracy rate of 88.4% (0.884). It was followed by the 12-variable (RA, NPM, FDR, ROIC, EM, LR, CR, ITR, ATR, GPM, ROI, CR) RF algorithm and the 10-variable (RA, NPM, FDR, ROIC, EM, LR, CR, ITR, ATR, GPM) KNN algorithm. The RF and KNN algorithms showed the same performance with an estimation accuracy rate of 81.1%. The SVM, DT, and NB algorithms yielded close values with estimation accuracy values of 74.7%, 73.7%, and 73.7%, respectively. The best accuracy values with SVM, DT, and NB algorithms were obtained in the 5, 11, and 6-variable estimation models, respectively. The DT and NB algorithms showed the lowest performances.

According to the performance metrics of precision, recall, F1-Score, and Cohen’s Kappa values, the most successful machine learning method was found to be the XGBoost algorithm.

**5. Conclusion**

In this study, the audit firm preferences of enterprises traded in Turkiye Borsa Istanbul Star market in sectors such as packaging, chemicals and plastic, metal goods and machinery, textile clothing and leather, transportation, manufacturing, base metal, defense, information and software, paper products, stone soil cement, furniture, and wholesale and retail trade were estimated through machine learning algorithms.

In these estimations, machine learning algorithms such as SVM, RF, DT, KNN, NB, and XGBoost were used. The purpose here is to establish an estimation model that predicts the relationship between dependent and input variables to have the best fit. While the dependent variable in the data sets of the models was the audit firm preference, the input variables were determined as financial ratios (RA, NPM, FDR, ROIC, EM, LR, CR, ITR, ATR, GPM, ROI, CR, LR). Among the 13 attributes used in the study, the highest prediction value was obtained by using the first 8 attributes, respectively. In selecting the attributes, the Mutual Information (MI) method was applied to find importance values. The ranking of importance of the attributes was determined as follows: RA, NPM, FDR, ROIC, EM, LR, CR, and ITR. It was determined that RA (Return on Assets) data was the most important attribute.

66 estimation models were established with each of the SVM, RF, DT, KNN, NB, and XGBoost machine learning methods for 11 different variable groups. In order to measure the success rate of the estimation models, the performance values used for classification such as accuracy, precision, recall, f1-score, and Cohen’s Kappa values were calculated. When ranked according to performance metrics, the best performance was displayed by the 8-variable (RA, NPM, FDR, ROIC, EM, LR, CR, ITR) XGBoost algorithm with an accuracy rate of 88.4%.

With these and similar artificial intelligence-supported estimation studies, it is thought that public companies will be able to make accurate estimations regarding their audit company preferences. In future studies, it is aimed to carry out estimation studies with different machine learning methods by creating larger data sets.