PressAcademia
www.pressacademia.org/journals/jefa
JEFA
ISSN 2148-6697
Journal of Economics, Finance and Accounting YEAR 2023 VOLUME 10 ISSUE 4

PREDICTING FINANCIAL DISTRESS USING SUPERVISED MACHINE LEARNING ALGORITHMS: AN APPLICATION ON BORSA ISTANBUL

DOI: 10.17261/Pressacademia.2023.1828
JEFA- V.10-ISS.4-2023(3)-p.217-223

Seyfullah Selimefendigil
Turkish-German University, Department of Business Administration Istanbul, Turkiye.
selimefendigil@tau.edu.tr, ORCID: 0000-0001-7017-9673

Date Received: October, 29, 2023 **Date Accepted:** December 21, 2023

To cite this document
Selimefendigil, S., (2023). Predicting financial distress using supervised machine learning algorithms: an application on Borsa Istanbul. Journal of Economics, Finance and Accounting (JEFA), 10(4), 217-223.
Permanent link to this document: http://doi.org/10.17261/Pressacademia.2023.1828

ABSTRACT

Purpose- The main purpose of this study is to identify the most significant variables to detect financial distress earlier and to find the best machine learning algorithm model.
Methodology-This study has used Support Vector Machine, Logistic Regression, Random Forest and K-nearest neighbors method techniques to predict the financial distress prediction for the companies of Turkey between 2012 and 2021.
Findings- As a result of the study, it has been determined that Random Forest provides the best results in terms of precision, accuracy, and recall. Further, this study has found the most important five independent variables to determine the financial distress status of the firms. In this way, it has been found that Current Assets/ Current Liabilities, Working Capital / Total Assets, Gross profit / Revenue, Retained Earnings / Total Assets and Sales growth rate are the most useful variables to determine financial distress status of Turkish firms earlier.
Conclusion- This study has concluded that cash ratios and profitability ratios and sales growth are the most important independent variables to determine financial distress one-year ahead. Furthermore, it has been found that random forest is the best machine learning method among other supervised machine learning methods used in this study.

Keywords: Financial distress, support vector machine, logistic regression, random forest, k-nearest neighbors
JEL Codes: G32, G33, C52.

1. INTRODUCTION

The prediction of bankruptcy is one of the most pressing issues in finance. As a result, financial distress (i.e. bankruptcy likelihood) prediction continues to be a hot topic in finance research (Elhoseny et al., 2022). Studies on predicting financial distress have been in progress for more than a half century. To identify the corporate solvency the financial distress prediction is a key issue. The primary objective of the financial distress prediction is to distinguish the stabilize companies from firms at the risk of financial distress. Financial risk is important for the investors as they decide to invest with their risk preferences. Regulators also benefit from the rapid identification of risk of each firm and are able to perform well in terms of supervision and management. The result of this has been a growing interest in the accurate prediction of business risks both in academia and in the business community (Qian et al., 2022).

While a consensus definition of financial distress remains elusive, it is acknowledged that varying degrees of financial distress exist. In its mildest form, financial distress may manifest as a shortage of cash. On the other hand, the most severe cases may involve a liquidity crisis or even bankruptcy (Özparlak and Özdemir Dilidüzgün, 2022). Although bankruptcy and financial failure are used interchangeably, bankruptcy is defined as the last resort to recover from a financial failure (Kinay, 2010).

Samuel & Gabel (1959) introduced the term "Machine Learning" and described it as a method of self-learning for computers without the use of a guide. Machine learning models build their models based on past data and improve their learning level independently (Gerçek & Özdemir Dilidüzgün, 2022). Machine learning techniques are regarded as the most popular algorithm techniques nowadays. These techniques are known for their accurate predictability performance. In case the outcome of the data is given previously then the supervised techniques are utilized. Furthermore, the supervised learning techniques categorize the outcome based on their labels (Özlem & Tan, 2022).

Recent global economic recession, highly volatile exchange rates, and a soured inflation rate have led to many firms in Turkey declaring bankruptcy (Aker and Karavardar, 2023). As of 2020, Turkey is the country with the highest number of bankruptcies and the second highest debt ratio among developing countries (Institute of International Finance, 2021). It has also been reported that 80 percent of newly established Turkish companies go into bankruptcy within their first five years of operation (Bloomberg, 2018). Accordingly, this study analyzes the bankruptcy likelihood of Turkish companies using supervised machine learning techniques. Based on supervised machine learning algorithms, this study has examined 477 companies that operate on the Borsa Istanbul exchange between 2012 and 2021. For this study, companies with negative net income for two consecutive years are defined as distressed and non-distressed otherwise.

This paper is divided into several sections. A discussion of national and international research is presented in the second part of this paper. The third section provides a description of how machine learning algorithms work as well as their methodology. The fourth section identifies the source of data and defines dependent and independent variables; the fifth section discusses results. As part of the final section of this study, the implications and limitations of the study are discussed.

2. LITERATURE REVIEW

The first dominant studies in financial distress prediction were conducted by Altman (1968), Ohlson (1980) and Zmijewski (1984). While each prediction model used different variables and statistical methods all models used accounting variables as a common feature (Avenhuis, 2013). Later on, several models were developed to predict financial distress of firms.

The study conducted by Oribel & Hanggraeni (2021) used Indonesian companies to determine their distress level. In their comprehensive study, the Support Vector Machine (SVM) method was applied, and it has ended up with 90% accuracy rate. This study followed Altman et al.'s (2010) definition of determining the financial distress of companies. Further it has been concluded that the linear SVM outperforms the radial and polynomial SVM models.

In their study, Qian et al. (2022) classified companies into distressed and non-distressed entities. Furthermore, they utilized a variety of machine learning methodologies including SVMs, artificial neural networks (ANNs), decision trees (DTs), random forests (RFs), and logistic regression in order to analyze the data. Accordingly, they found that the gradient boosted decision tree with the corrected feature selection measure outperforms all other models.

In a more extended context Elhoseny et al. (2022) have examined financial distress and credit risk assessment. They use companies from Taiwan, Australia and Poland as a sample to determine their financial distress and credit risk assessment. In this way, a novel approach has been developed and put forward. In terms of accuracy and precision, the adaptive whale optimization algorithm (AWOA-DL) has been compared with other models, including DNN, TLBO-DL, LR, and RBF Network. According to their results, this novel approach allows for more precise fine-tuning parameters and achieved a 95.8% accuracy rate with its dominance compared to other methods. To build on this, Tsai et al. (2014) used three different machine learning techniques to predict the bankruptcy likelihood of German, Australian and Japanese firms. In this way, multilayer perceptron (MLP) neural networks and SVM are compared with decision trees with a boosting method. Consequently, it has been found that decision trees with a boosting method provide higher accuracy.

Lin et al. (2011) have selected a few features to conduct machine learning techniques. A number of important variables have been identified through the use of data mining techniques. A total of 74 financial ratios have been selected as the best subset of the variables of companies listed on the Taiwan Stock Exchange. Afterward, 5 selected ratios were used to predict financial distress for firms one year ahead. A comparison was made with other classic models (Altman Edward I., 1968; Beaver, 1966; Zmijewski, 1984; Ohlson, 1980). As a result, the model with selected features outperforms classical models, and this was conducted using MDA, Logit, Neural Network, and SVM models. Additionally, the SVM model produces better results when certain variables are considered.

By focusing on the Turkish context, it is evident that several different studies have applied machine learning algorithms to predict financial distress. A recent study conducted by Aker and Karavardar (2023) has used Logistic Regression, Decision Tree, Random Forest, Support Vector Machines, K-Nearest Neighbor and Naive Bayes algorithms to predict financial distress of Turkish firms earlier. They have found that Naive Bayes has a superior prediction ability than other models. In a similar vein, İçerli (2005) has examined the financial distress of Turkish firms for the years between 1990 and 2003. In comparison to other algorithms they use, such as logistic regression and discriminant analysis, artificial neural networks are better at predicting financial distress. To build on this Aksoy and Boztosun (2018) investigated the same prediction by using manufacturing firms operating in Turkey between 2006 and 2009. Using multiple discriminant analysis and logistic regression, they concluded that logistic regression is more effective at detecting financial distress early.

3. METHODOLOGY

3.1. Logistic Regression

Logistic regression is considered as a classification method rather than regression model and thus it transfers the probability value into 0 or 1. In logistic regression the outcome variable takes two different variables i.e. binary. This method is very useful to predict categorical variables. The binary outcome is estimated with the independent variables to acquire information. Logistic regression employs maximum likelihood of observing data.

$$\text{logit}(p)=\ln(p/1-p) = \alpha+ \beta1X1+\ldots. BnXn \quad (1)$$

In this equation increasing X by one unit changes the log odds by β1. In addition to that, independent variables shouldn't be multicollinear and the log-odds of the outcome and independent variables should be linear.

3.2. Support Vector Machines

Based on statistical learning theory, Support Vector Machines are machine learning algorithms. By utilizing feature function fitting, this method is able to work with samples of small, non-linear data for high dimensional pattern recognition (Cortes & Vladimir, 1995). SVM is used to separate two different classes or to detect the outliers. This method is especially handy where two different classes exist. In this manner, SVM uses hyperplane to categorize the variables. This yield better results compared to other methods in terms of classification (Malik et al., 2021). The original SVM algorithm can be expressed as mathematical formula below:

$$y(x)=sign\ (\Sigma yi\ \alpha iK(x,xi)+b) \quad (2)$$

Where y(x) represents the predicted class label for the input vector x; y_i is the class label for the i-th example; α_i are the Lagrange multipliers obtained during training; $K(x_i, x)$ is the radial kernel function that measures the similarity between two feature vectors; b is the bias term. In this sense there are different kernels to be chosen in SVM algorithms such as linear, kernel and polynomial. The linear kernel is applied where the model can be classified by a linear decision boundary whereas radial kernel is applied for the datasets which is suitable for most of the dataset and has versatile functions.

3.3. Random Forest

Random forest techniques belong to ensemble learning family and used for classification and regression tasks. This technique is known for its robustness and accuracy to handle high dimensional datasets. In this technique, multiple decision trees are combined and used to predict an outcome of the model. Each decision tree in the random forest is constructed by recursively dividing the feature space. The splitting process has a target of gaining more information at each node (Breiman, 2001).

The random forest is also well known for its variable importance measures. In this way, this technique has two different methods for measuring variable importance, namely mean decrease Gini and mean decrease accuracy. While the former one is the decrease in Gini impurities for the predictor across the forest, the latter one is the average decrease in accuracy for the predictor after permuting (Nicodemus, 2011).

3.4. K-nearest Neighbors (KNN) Method

K-nearest neighbors (KNN) method has both classification and regression features. First KNN algorithms leave a distance between observed data and further identify new data with not known target. In this learning method either Euclidean or Manhattan distances are employed to measure the proximity between variables. The K parameter in this model is used to determine the number of neighbours for this model. The optimal K parameter is chosen based on the cross-validation techniques. During the training KNN stores independent variables vectors and their corresponding values. During the prediction the distance between query point is calculated and the K nearest neighbour is selected based on the majority or average value (Zhang, 2016).

4. DATA

This study uses the several data from Turkey. In this manner the financial data of 227 firms listed in Borsa İstanbul from 2012 and 2021 has been extracted from Thomson-Reuters database. To label the target variables the firms are labelled as D (distressed) and ND (non- distressed) to represent their financial distress status. To determine the distress status of companies this study considered their net income. Following previous literature, companies with negative net income for two consecutive years are classified as distressed and non-distressed otherwise (Altman Edward I., 1968; DeAngelo & DeAngelo, 1990; Hill et al., 1996; Li & Sun, 2008; Oz & Yelkenci, 2017; Oz & Simga-Mugan, 2018).

By combining the sample and floor functions, the dataset has been divided into training and test sets. Data representing 80% of the dataset is used as training data and data representing the remaining 20% of the dataset is used as test data. A ratio of 8:2 is used for the distribution of the test and training sets, in accordance with previous studies (Oribel & Hanggraeni, 2021; Lin et al., 2011). This study contains 27 independent variables, and the objective is to identify the five most significant variables from them. Appendix-A contains a list of all independent variables that were used. Accordingly, the variable importance measure has been implemented, and these variables are measured in descending order. By ensuring the absence of multicollinearity, the five most important independent variables were selected for logistic regression analysis. The variables identified as the most useful variables and the variables used in other studies were found to be consistent (see: Altman Edward I., 1968; Zmijewski, 1984 ;Ohlson, 1980). The selected variables are listed in Table 1, along with their formulas, and descriptive statistics are shown in Table 2.

Table 1: Independent Variables

No	Formula
V1	Current Assets/ Current Liabilities
V2	Working Capital / Total Assets
V3	Gross profit / Revenue
V4	Retained Earnings / Total Assets
V5	Sales growth

Table 2: Descriptive Statistics

Variable	Minimum	Maximum	Mean	Median
V1	0.02388	136.58753	1.85775	0.96466
V2	3.40526	0.99779	0.13798	0.12590
V3	-1.2907	1.0349	0.2385	0.2099
V4	-51.409	0.40127	-0.01571	0.02066
V5	-17.567	57.35006	0.03025	-0.04180

5. RESULTS AND DISCUSSIONS

Table 3 below presents the results of the logistic regression after labeling the outcome variables as D and ND. To assess the multicollinearity among independent variables, the variance inflation factor was applied. Table 4 illustrates the variance inflation factor for the variables.

Table 3: Logistic Regression Results

Variable	Estimate	Std. Error	z value Pr(>|z|)	Pr(>|z|)
Intercept	0.84096	0.16556	5.080	3.78e-07***
V1	0.16307	0.13301	1.226	0.220
V2	2.40802	0.43395	5.549	2.87e-08***
V3	2.39252	0.49560	4.827	1.38e-06***
V4	0.02005	0.05166	0.388	0.698
V5	-0.01838	0.03863	-0.476	0.634

Table 4: Variance Inflation Factor

Variable	Estimate
V1	1.603506
V2	1.605755
V3	1.001814
V4	1.001963
V5	1.002182

On the basis of the above test results, it can be concluded that the gross profit percentage ratio and the ratio of working capital divided by total assets are important indicators of financial distress one year in advance. The variance inflation factor levels show that there is no need to concern multicollinearity in this dataset.

In Table 5, the precision, accuracy, and recall values of other machine learning techniques such as KNN, SVM, and Random Forest methods have been presented. These techniques are evaluated based on the precision, accuracy and recall metrics. These metrics are calculated based on the true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

- Precision: Precision is the ratio of correctly identified positive cases in a classification scenario. This metric is computed as proportion of true positive predictions on the total positive predictions.
- Accuracy: As a percentage, accuracy measures the percentage of cases correctly identified to the total number of cases.
- Recall: Recall measures the proportion of positive cases that are correctly estimated to all positive cases

Table 5: Classification Performance Comparison

Models	Precision	Accuracy	Recall
SVM	0.84	0.88	0.94
Random forest	0.91	0.95	0.99
KNN	0.72	0.79	0.95
Formula	(TP /TP+ FP)	(TP +TN/ TP+ TN+ FP+ FN)	(TP/ TP+FN)

According to test results it can be concluded that random forests outperform other methods in terms of accuracy, precision and sensitivity. This method is followed by the SVM model which has been conducted with linear kernel. The last method, namely KNN, is less accurate compared to other methods. The test result has shown that the financial status of a firm can be predicted one year ahead whether the firm is in a distressed position or not.

6. CONCLUSION

Predicting financial distress is an important component of risk management, especially in countries with high inflation, such as Turkey. Detecting financial distress early can prevent creditors from incurring losses. Additionally, this early detection mechanism will help to mitigate the impact of bankruptcy on shareholders, employees, and other stakeholders. In order for a country's economy to be in good shape, companies must operate efficiently and without difficulty. This implies that the consequences for the financial health of firms do not just affect microeconomics, but also macroeconomics.

In this study, early financial distress detection has been measured through several supervised machine learning models. In this way, 227 firms have been used from Borsa Istanbul between the years of 2012 and 2021. As a result, it has been concluded that cash ratios and profitability ratios and sales growth are the most important independent variables to determine financial distress one-year ahead. Furthermore, it has been found that random forest is the best machine learning method among other supervised machine learning methods used in this study. It may be beneficial for firms that feel likely to go bankrupt to focus on the most important factors that will enable them to recover sooner, or to avoid going bankrupt. Several implications are derived from the findings of this study for policy makers, managers, and academics alike.

Despite its strengths, this study is not without limitations. First, the selected variables and logistic regression results cannot be generalized to all countries. As a result of the limited number of methods used, the results of deep learning, neural networks, etc. methods have not been evaluated. To evaluate the predictive ability of the independent variables in this study, further studies should consider other methods. Financial distress has been predicted solely through financial variables in this study; however, other non-financial metrics (for example, the number of employees, the existence of an audit committee, board composition, firm age) and macroeconomic variables (for example, inflation rate, exchange rate, interest rate) should also be considered to arrive at new insights. Moreover, the impact of the recent financial crisis COVID-19 can be incorporated in order to determine how it plays a moderating role in the emergence of financial distress indicators.

REFERENCES

Aker, Y., & Karavardar, A. (2023). Using machine learning methods in financial distress prediction: sample of small and medium sized enterprises operating in Turkey. Ege Akademik Bakis, 23(2), 145-162

Aksoy, B., & Boztosun, D. (2018). Financial failure prediction by using discriminant and logistics regression methods: evidence from BIST manufacturing sector. Finans Politik & Ekonomik Yorumlar, 646, 9–32.

Zhang, L., Altman, E. I. & Yen, J. (2010). Corporate financial distress diagnosis model and application in credit rating for listing firms in China. Frontiers of Computer Science in China, 4, 220-236.

Altman Edward I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 23(4), 189–209.

Oude Avenhuis, J. (2013). Testing the generalizability of the bankruptcy prediction models of Altman, Ohlson and Zmijewski for Dutch listed and

large non-listed firms (Master's thesis, University of Twente).

Beaver, W. H. (1966). Financial ratios as predictors of failure. Journal of Accounting Research, 71-111.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273-297.

DeAngelo, H., & DeAngelo, L. (1990). Dividend policy and financial distress: an empirical investigation of troubled NYSE firms. Journal of Finance, 45(5), 1415-1431.

Elhoseny, M., Metawa, N., Sztano, G., & El-Hasnony, I. M. (2022). Deep learning-based model for financial distress prediction. Annals of Operations Research, 11(2), 1-23.

Hill, N. T., Perry, S. E., & Andes, S. (1996). Evaluating firms in financial distress: An event history analysis. Journal of Applied Business Research (JABR), 12(3), 60-71.

İçerli, M. Y. (2005). Prediction of Financial Failure in Businesses and an Application (PhD thesis, University of Dokuz Eylül)

Institute of International Finance. (2021). Global Debt Monitor COVID Drives Debt Surge — Stabilization Ahead ?

Kinay, B. (2010). Ordered Logit Model approach for the determination of financial distress. Numéro spécial, 119-131.

Li, H., & Sun, J. (2008). Ranking-order case-based reasoning for financial distress prediction. Knowledge-Based Systems, 21(8), 868-878.

Lin, F., Liang, D., & Chen, E. (2011). Financial ratio selection for business crisis prediction. Expert Systems with Applications, 38(12), 15094-15102.

Malik, H., Fatema, N., & Iqbal, A. (2021). Intelligent data-analytics for condition monitoring: smart grid applications. Academic Press.

Nicodemus, K. K. (2011). On the stability and ranking of predictors from random forest variable importance measures. Briefings in Bioinformatics, 12(4), 369-373.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 5(2), 109-131.

Oribel, T., & Hanggraeni, D. (2021). An application of machine learning in financial distress prediction cases in Indonesia. International Journal of Business and Technology Management, 3(2), 98-110.

Oz, I. O., & Simga-Mugan, C. (2018). Bankruptcy prediction models' generalizability: Evidence from emerging market economies. Advances in Accounting, 41, 114-125.

Oz, I. O., & Yelkenci, T. (2017). A theoretical approach to financial distress prediction modeling. Managerial Finance, 43(2), 212-230.

Özlem, Ş., & Tan, O. F. (2022). Predicting cash holdings using supervised machine learning algorithms. Financial Innovation, 8(1), 1-19.

Özparlak, G., & Dilidüzgün, M. Ö. (2022). Corporate bankruptcy prediction using machine learning methods: the case of the USA. Uluslararası Yönetim İktisat ve İşletme Dergisi, 18(4), 1007-1031.

Qian, H., Wang, B., Yuan, M., Gao, S., & Song, Y. (2022). Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree. Expert Systems with Applications, 190, 116202.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3), 210-229.

Tsai, C. F., Hsu, Y. F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. Applied Soft Computing, 24, 977-984.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of Translational Medicine, 4(11), 213-229.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research, 3(4), 59-82.

APPENDIX A: INDEPENDENT VARIABLES

BETA	PRICE TO BOOK RATIO	CASH TO TOTAL ASSETS
RETURN ON ASSET	QUICK RATIO	ASSETS GROWTH RATE
RETAİNED EARNINGS TO TOTAL ASSETS	WORKING CAPITAL TO REVENUE	ACCOUNT RECEIVABLES TURNOVER

OPERATINGMARGIN RATIO	EQUITY TO TOTAL ASSETS	GROSSMARGINRATE
CURRENT RATIO	SALES GROWTH RATE	ACCOUNTSPAYABLETURNOVER
WORKING CAPITAL TO TOTAL ASSETS	CASH TO REVENUE	REVENUE TO COST OF GOODS SOLD
WORKING CAPITAL TO TOTAL LIABILITIES	CASH TO TOTAL LIABILITIES	REVENUE TO TOTAL ASSETS
MARKT VALUE OF EQUITY TO TOTAL LIABILITIES	BOOK VALUE PER SHARE	INVERTORY TO CURRENT ASSETS
OPERATING EXPENSES TO TOTAL ASSETS	EARNIGS PER SHARE	REVENUE TO EQUITY