Adv. Math. Fin. App., 2024, **9**(3), P. 856- 872

Advances in Mathematical Finance & Applications
www.amfa.iau-arak.ac.ir
Print ISSN: 2538-5569
Online ISSN: 2645-4610

Original Research

Designing a model for predicting corporate bankruptcy using Ensemble learning techniques

Hosseine Egbbali[a], Alimohamad Ahmadvand[a] *

[a]*Department of Industrial Engineering, University of Eyvanekey, Eyvanekey, Iran*

ARTICLE INFO

Article history:
Received 2023-01-12
Accepted 2024-01-23

Keywords:
Bankruptcy
Ensemble learning techniques
Prediction
Stacking Method

ABSTRACT

The bankruptcy of corporations causes huge losses for investors, managers, creditors, employees, suppliers, and customers. If someone understands the reason for the corporate's bankruptcy, then he can save the corporate from certain death with the necessary planning. Therefore, bankruptcy forecasting is the most important prerequisite for bankruptcy prevention. Due to this issue, the main aim of this article is the prediction of the economic bankruptcy of corporations in the Tehran Stock Exchange using group machine learning algorithms. Financial ratios have been used as independent variables and healthy and bankrupt corporations as research dependent variables. The statistical population of the study is the information of financial statements of corporations on the Tehran Stock Exchange from the years 2004 to 2021. In this study, sampling is not used and corporations include two groups healthy and bankrupt. The bankrupt and non-bankrupt groups are selected based on the threshold of the Springate model. The research findings indicate that the accuracy of predicting the bankruptcy of corporations in the group learning model by stacking method is higher than other used models where the AUC and Accuracy Ratio were 0.9276 and 0.8247, respectively.

1 Introduction

Generally speaking, bankruptcy is a part of the foundation and natural component that modern economies are built upon [1]. It is not entirely accurate to say that bankruptcy is a natural component or foundation that modern economies are built upon. While bankruptcy is a legal process that allows companies and individuals to restructure their debts or liquidate their assets when they are unable to pay their creditors, it is not a desirable outcome for any economy. In fact, bankruptcy is often seen as a negative outcome because it represents a failure of a business or individual to meet their financial obligations. Bankruptcy can lead to job losses, creditor losses, and economic instability, which can have negative consequences for the broader economy [2]. The statement is generally true. Misclassifying firms can have significant financial consequences, particularly for bankrupt firms. If a firm is misclassified as not being at risk of bankruptcy when it actually is, then lenders and investors may continue to

* Corresponding author. Tel.: *09012797836*
E-mail address: *alimohammadahmadvand@yahoo.com*

provide credit and investment, which can result in significant losses if the firm eventually goes bankrupt. On the other hand, if a firm is misclassified as being at risk of bankruptcy when it is actually financially stable, then the firm may face difficulties in obtaining credit and investment, which can harm its operations and growth potential. Therefore, accurate bankruptcy prediction models can be very valuable for lenders, investors, and other stakeholders in the economy [3]. By using these models to identify firms that are at risk of bankruptcy, they can take appropriate measures to reduce their exposure to risk, such as reducing or withdrawing credit and investment. This can help to prevent losses and stabilize the economy [4]. Furthermore, the cost of misclassifying bankrupt firms can be particularly high, as these firms may have significant outstanding debts and liabilities that need to be resolved. Misclassifying a bankrupt firm as financially stable can result in significant losses for creditors and investors, as well as the potential for lengthy legal proceedings to recover their losses. Therefore, the development and expansion of accurate and well-performing models for bankruptcy prediction is important to help reduce the potential financial costs of misclassifying firms, particularly for bankrupt firms [5].During the past few years, novel approaches to machine learning techniques containing powerful data manipulation tools have been introduced which consider on a larger scale as well [4]. This issue can enhance the possibility of developing more advanced and new models. Therefore, this gives the opportunity to users to not only process more data but also not limit models due to linearity and variable selection. To the best of our knowledge and what has been reported in the open literature, there is no research study at the moment in which machine learning techniques have been tested on the management, sector, and financial statements for predicting bankruptcy [6]. Therefore, it is believed that there is great potential to develop more accurate and well-performing models, employing these approaches on a large dataset with a significant number of characteristics. The existing models are optimized to perfectly classify bankrupt firms at the expense of lower overall precision. This issue can be caused due to the high cost associated with incorrectly classifying the corporate that is actually going bankrupt. Hence, the present models are completely aligned with the point of view of the investors, banks, rating agencies, and shareholders. The obtained results from the present research study can give further life to our faith that management, sector, and financial statements are appropriate and verified data that can be used for the prediction of bankruptcy. Additionally, more complicated mathematical models such as random forest (RF) and artificial neural networks (ANN) demonstrate great potential in the prediction of bankruptcy, being able to find and handle patterns in a large dataset [7].Credit analysis and bankruptcy predictions have existed for the past decades. The earliest evidence goes back to the 1890s and is associated with the likelihood estimation [8]. Privately-owned banks mainly utilized the noted analysis to grant loans to firms according to their creditworthiness, therefore, this spreads the plan of ratio analysis. In the early 1900s, many attempts were made to standardize the structure which eventually helped to increase the credit men [8]. Later on, in 1919, the 1st ratio analysis of the federal bank in the United States was reported by the Federal Reserve to commence a public discussion on credit risk and to gain public momentum [8]. In the field of bankruptcy prediction, Beaver [9] can be introduced as one of the earliest forerunners. A univariate analysis was used by Beaver to figure out considerable differences in several specified variables for two categorical groups of non-bankrupt and bankrupt corporations. It should be mentioned that the performed analysis was carried out on a sample of 706 firms for 5 years. The sample was chosen to keep out some specified sectors, and the division between the non-bankrupt and bankrupt corporations was just about fifty percent for all the years. Thirty picked variables were also divided into five various subgroups arranged by characteristics. The noted subgroups were associated with the various sections of the firms' financial structure, like cash flows, ratios related to acid tests, turnover, and

net income. According to the obtained model, four different propositions were created for the identification of the distressed firms and appropriate thresholds for each of mentioned ratios. At the moment, the suggested thresholds are introduced and popular among scholars as rules of thumb for the aforementioned ratios. Moreover, in 1968, the trend of utilizing financial data was started by Beaver [10] to systematically rate firms by creditworthiness. In the same year, Beaver also introduced how investors view distress as well as alternative ratios from the stock market point of view [11]. Another framework to predict bankruptcy was proposed by Boyacioglu et al. [12] in 2009. In the noted research study, a ratio model was created for Turkish banks from 1988 to 2000. The main purpose to create this model was the prediction of future bank failures right after the financial crisis of 2007 and 2008. To this end, for the base case, 20 various financial ratios with 6 different characteristic groups containing management quality, asset quality, capital adequacy, liquidity, sensitivity to market risk (CAMELS), and earnings were chosen as predictor variables. With the main idea of improving the prediction performance, 4 different datasets including 22 failed and 44 non-failed banks with various features were developed. Four different approaches called learning vector quantization, competitive learning, multi-layer perceptron, and self-organizing map were utilized in the category of neural networks. The obtained results demonstrated that learning vector quantization and multi-layer perceptron were the most accurate and successful models to predict the financial failure of banks. The Bloomberg DRSK models have been recently introduced as another well-known bankruptcy predictor within finance [13, 14]. It is noteworthy to mention that in contrast to the formerly proposed models, the Bloomberg DRSK models are more comparable to credit modeling. Besides, the Bloomberg DRSK models were established based on Merton's Distance model. It was suggested for credit modeling in the beginning and subsequently, it was developed to be incorporated in the popular Black-Scholes model for option pricing. Frictionless trading, short selling, continuous trading, and the prices following the Brownian motion can be mentioned as the underlying assumptions of the problem [15]. The created models were unfocused and focused models, whereas the latter was purposed for financial firms [13, 14]. Furthermore, the observations were also separated based on the size, leading to the creation of four mutually collectively and exclusive exhaustive models. Depending on the out-of-sample years, an accuracy of 85.6% to 87.8% was obtained for the model proposed for non-financial private corporations. The main privilege of the Bloomberg DRSK models is the ecosystem that the models are applied in, called the Bloomberg terminal. Bloomberg terminal contains the most recent financial data existing with trailing financial statements. Additionally, Bloomberg suggests reclassified financial statements for large US corporations as well. This leads to improving the truthfulness of the financial statement as well as enhancing accuracy. The suggested models [13, 14] were established on the same assumptions as the models by Zhang et al. [16] and Crouhy et al. [17].

Furthermore, Næss et al. [18] completed the latest addition to the Norwegian markets based on Wahlstrøm and Helland's research study [19]. The mentioned study could harmonize predictions between the Norwegian-developed SEBRA model, the famous Altman Z-score, and their own sets of variables. The models were carried out on a range of statistical techniques for each ratio set. Statistical techniques were GAM, SVM, LDA, CT, GLM, QDA, NN, and KNN. For the NN method, in addition to an approach involving reduction of the dimensionality, both backward and forward sweeps of back-propagation were implemented. Similar samples collected from corporations with financial statements from 2005 to 2014 were used to train and test each model. the logic of Bernhardsen and Larsen in Ref. [20] was followed by Wahlstrom and Helland [19] in their study in which corporations with low total assets (< 500 TNOK) were excluded. Moreover, in order to be consistent with Bernhardsen and Larsen's

research study [20] and make comparison, financial firms were also excluded in Ref. [19]. By following the logic proposed by Boyacioglu et al. [12], the distribution between the non-bankrupt corporations and bankrupt corporations was manipulated. The dataset was divided into 1/3 bankrupt corporations and 2/3 non-bankrupt corporations in which a random option was selected for the sampling of non-bankrupt corporations. On the other hand, all of the bankrupt corporations were included. This was in contrast to Berg's study in 2007 [21]. In that research, a true distribution was utilized and a probability threshold was adjusted for classification. To create a model, it was tried to scale the financial statement to a mean of zero in order to eliminate the possibility that statistical techniques augment significance to size, rather than distance and distribution [19]. To obtain the best predictions, the hyperparameters of the models were tuned. Besides, to ensure that the obtained results have reproducibility and were not a product of random chance, the models were cross-validated. It is not an easy task to determine the exact reason or reasons for bankruptcy in any particular case. In many cases, several reasons together may lead to bankruptcy. The analysis of financial statements requires tools and techniques that enable analysts to examine current and past financial statements aimed at evaluating the performance and financial status of a corporate and estimating the possibility of future and potential risks. In this research, we proposed a model for predicting the financial bankruptcy risk of listed and OTC corporations by utilizing Ensemble learning algorithms [8]. New methods utilize machine learning, including ensemble learning techniques like bagging and boosting. The main advantage of the new machine learning methods is they provide more accurate predictions of company bankruptcy compared to previous qualitative methods. The increased prediction accuracy stems from the machine learning models' ability to capture complex patterns and relationships in data that may be hard to discern. So the main gap or difference between the old and new methods is the higher level of prediction accuracy achieved through leveraging machine learning algorithms and techniques.

Data-driven machine learning models can process much larger sets of financial, operational, and macroeconomic data compared to limited qualitative assessments [7]. This enables the models to uncover more predictive insights. Ensemble techniques like random forests, gradient boosting, and stacking generalization allow for combining multiple predictive models to improve overall accuracy. This is more systematic versus subjective qualitative opinions. Machine learning can continually update predictions in real-time as new data comes in, while qualitative methods provide episodic and periodic assessments of bankruptcy risk. In summary, the automation, sophistication, and customization of modern machine learning makes it significantly more accurate than previous subjective and limited qualitative techniques for predicting company bankruptcy. The gap in accuracy continues to widen as machine learning capabilities improve [4].

In the second part, we will have a general reference to the research methodology. Then, in more detail, we first state the financial ratios for predicting the bankruptcy of companies and select those financial ratios that have the most power to predict bankruptcy using the WOE technique. After that, we define the basic prediction models. By implementing 3 basic models on the database, we get the results of each along with the accuracy of that method. Then we will implement the proposed research model, collective learning, on the results of the basic models and compare the accuracy of our method with previous similar articles.

2 Methodology

Choosing a research methodology is one of the most important researcher's efforts since it should be determined what approach and method to adopt to help get the answers to the research questions as

accurately, easily, and quickly as possible. It depends on the nature, objective, and subject matter. The research method is the way to achieve the goal and guarantees the success of the research. The research method contains a set of measures to identify the truth and achieve the designed goals. To perform this research, all corporations listed on the Tehran Stock Exchange from March 2004 to March 2021 were considered as the statistical population and the statistical sample was extracted from these corporations. The total number of year-corporate is 1652. Since the data required is in the form of year-corporate and each corporate may have been studied in a few years (from 2004 to 2021), thus, the total number of data accounted for 16984 year-corporate. The solution method in this research was set to first use the combined machine learning methods to predict the financial bankruptcy risk of listed corporations and over-the-counter corporations on the Tehran Stock Exchange. We utilized the ensemble learning method, which is one of the new areas of machine learning. Ensemble learning is a field of machine learning in which, instead of using a model to solve a problem, employs several models in combination to increase the model's output estimation power. We used the stacking technique in this research, which is one of the strongest ensemble learning techniques. In the first stage, 3 basic forecasting models, including logistic regression model, SVM model, and gradient boosting, were trained by the data. In the second stage, the final model of the education system was trained that plays the role of the final decision maker and learns in the process of learning how each model works. Due to the power of the basic models, a weight is determined and assigned for each of them to use in the decision-making process, which is the very predicting financial bankruptcy. Using this ensemble learning technique, the predictive results would be more reliable.

2.1 Research Steps

2.1.1. Review of Predictor Variables Using Previous Research

Financial ratios have been used by analysts since 1870. At the moment, financial ratios analysis is an effective tool a and strong technique for users to assess the performance of future, present, and past, and to this day, owing to the development of science, technology, and knowledge in information and computing, there have been several improvements in using financial ratios [22]. In many research papers in the field of bankruptcy prediction, financial ratios were utilized for analysis and as a consequence, a background of all used financial ratios in the existing bankruptcy models was classified and documented in Table 1 [10, 11, 23-30].

2.1.2. Choosing the Main Variables

For selecting the main features (variables) in the current research paper, the weight of evidence (WOE) algorithm was utilized as a feature selection method out of 138 technical features that have been provided as input predictor features to the system, the features that enhance the precision of bankruptcy predictions. The WOE can be calculated by $ln\left(\frac{\%\text{events}}{\%\text{nonevents}}\right)$ for any feature. The predictive power of a single feature concerning its independent feature is told by WOE. In comparison to the proportion of non-events, if any of the bins/categories of a feature has a large proportion of events, a high value of WoE is gotten that in turn tells that the noted class of the feature separates the events from non-events. As mentioned above regarding the WoE value, it should be noted that the WoE value also says the predictive power of each bin of a feature. However, in the feature selection, a single value showing the whole feature's predictive power is useful. The equation for the information value (IV) is $\sum_{i=1}^{n}\left(\text{WOE}_i * (\%\text{events-}\%\text{nonevents})\right)$ where the IV is always a positive number.

Table 1: Financial Ratios

x1	Operating Profit	x70	Commercial and non-commercial accounts receivable
x2	Net other operating income and expenses	x71	Accounts and documents of commercial and non-commercial payables
x3	Financial costs	x72	Cost ratio
x4	Net cash inflow (outflow) of investment activities	x73	Working capital
x5	Cash balance at the beginning of the year	x74	Debt ratio
x6	Legal reserve	x75	Ownership ratio
x7	Tangible fixed assets	x76	Debt service
x8	Total liabilities and equity	x77	Current debt ratio
x9	Cash balance at the end of the year	x78	Financial to operating profit
x10	Inventories	x79	Accumulated profits to equity
x11	Taxes paid	x80	Accumulated profit to total debts
x12	Total debt	x81	Retained earnings on total assets
x13	Net cash inflows and outflows of operating cash	x82	Receipts to total assets
x14	Saved end-of-staff service	x83	Operating profit to current assets minus current liabilities
x15	Fund	x84	Net profit to total assets
x16	Net cash inflows and outflows for tax purposes	x85	Net profit to sell
x17	Equities	x86	Special interest before interest and tax on equity
x18	income tax	x87	Net profit to equity
x19	Intangible assets	x88	Operating profit to sell
x20	Gross profit	x89	Interest cost coverage ratio
x21	Total current liabilities	x90	Special interest on interest and taxes on current debt
x22	Sales, administrative and general expenses	x91	Frequency of inventory turnover
x23	Net cash inflows (outflows) from financing activities	x92	Inventory turnover period
x24	Net increase (decrease) in cash	x93	Sell to inventory
x25	Net profit	x94	Working capital to total debts
x26	Received financial facilities	x95	Working capital to total assets
x27	Total current assets	x96	Sales to total assets
x28	Total non-current liabilities	x97	Relative to the current
x29	Profit (loss) accumulated at the beginning of the year	x98	Instant ratio
x30	Profit before tax	x99	Inventory to working capital
x31	Net other non-operating income and expenses	x100	Current assets to total assets
x32	Orders and prepayments	x101	Financial costs to total debt
x33	Net cash inflows (outflows) from the return on investment activities	x102	Financial to net sales
x34	Accumulated profit (loss)	x103	Cash inventory to total assets

Table 1: continue

x35	Accounts and documents of commercial and non-commercial payables	x104	Short-term cash balance
x36	Definable profit	x105	Inventory to current assets
x37	Total Non-Current Assets	x106	Short-term operating costs
x38	Cash balance and fund bank	x107	Total operating costs
x39	Total assets	x108	Total costs to total sales
x40	Net sales or operating income	x109	Sale or return on fixed assets
x41	Cost of	x110	Sales or revenue to accounts receivable
x42	Annual adjustments	x111	Short-term sales or revenue
x43	Commercial and non-commercial accounts receivable	x112	Net sales to working capital
x44	Long-term financial facilities received	x113	Gross profit on sales or revenue
x45	Total receipts received	x114	Gross profit to total assets
x46	Average of total assets	x115	Gross profit for the short term
x47	Profit before tax on total assets	x116	Total Cash - Cash Inventory} Sales or Revenue}
x48	Profit before tax on total debt	x117	Sales or revenue to inventory
x49	Income or sales on total debt	x118	From interest and taxes to profits before the total
x50	Total current assets	x119	Net profit to warehouse
x51	Equity to fixed assets	x120	Equity-equity to total assets}
x52	Equity of long-term assets	x121	Working capital to fixed assets
x53	Accounts paid on sales revenue	x122	Fixed assets to total assets
x54	Cash flow sold or earned	x123	Receipts to total debts
x55	Cash flow to assets	x124	Financial cost to receive
x56	Cash flow to equity	x125	Financial costs to current debt
x57	Cash flow to total debt	x126	Cash flow to current debt
x58	Long-term amount to total assets	x127	Cash flow to working capital
x59	Short-term net sales facility	x128	Financial to net profit
x60	Short-term facility to current debt	x129	Net income logarithm
x61	Short-term debt to total expenses	x130	Logarithm of total assets
x62	Current debt to net sales	x131	Debt to banks
x63	Current liabilities to total assets	x132	Inventory of materials and goods * 365 sold or earned
x64	Current debt of special value	x133	Receivables * 365 Sold or earned
x65	Debtors or accounts receivable to total assets	x134	Short term * 365 sold or earned
x66	Gross profit plus sales or revenue	x135	Short term * 365 at the cost of goods sold
x67	Gross profit in addition to the benefit of total assets	x136	Operational operations to tangible assets
x68	Net profit plus interest expense on total assets	x137	Operations on intangible assets
x69	Current assets - Inventory Long-term assets}	x138	Inventories

Table 2: Fixed Rule IV

Information Value	Predictive power
<0.02	Useless
0.02 to 0.1	Weak predictors
0.1 to 0.3	Medium Predictors
0.3 to 0.5	Strong predictors
> 0.5	Suspicious

The final variables of this research (72 pcs) were extracted using the WOE method to predict the bankruptcy of corporations Index in Table 3:

Table 3: The Final Variables to Predict the Bankruptcy

x1	Operating Profit	x71	Accounts and documents of commercial and non-commercial payables
x2	Net other operating income and expenses	x72	Cost ratio
x3	financial costs	x73	Working capital
x6	Legal reserve	x77	Current debt ratio
x7	Tangible fixed assets	x9	Cash balance at the end of the year
x10	Inventories	x79	Accumulated profits to equity
x11	Taxes paid	x82	Receipts to total assets
x12	Total debt	x88	Operating profit to sell
x14	Saved end-of-staff service	x83	Operating profit to current assets minus current liabilities
x15	Fund	x84	Net profit to total assets
x16	Net cash inflows and outflows for tax purposes	x85	Net profit to sell
x17	Equities	x86	Special interest before interest and tax on equity
x18	income tax	x87	Net profit to equity
x22	Sales, administrative and general expenses	x91	Frequency of inventory turnover
x25	Net profit	x94	Working capital to total debts
x28	Total non-current liabilities	x97	Relative to the current
x29	Profit (loss) accumulated at the beginning of the year	x98	Instant ratio
x30	Profit before tax	x102	Financial to net sales
x31	Net other non-operating income and expenses	x103	Cash inventory to total assets
x32	Orders and prepayments	x104	Short-term cash balance
x33	Net cash inflows (outflows) from the return on investment activities	x110	Sales or revenue to accounts receivable
x34	Accumulated profit (loss)	x111	Short-term sales or revenue
x38	Cash balance and fund bank	x107	Total operating costs
x41	Cost of	x110	Sales or revenue to accounts receivable
x42	Annual adjustments	x111	Short-term sales or revenue
x43	Commercial and non-commercial accounts receivable	x112	Net sales to working capital
x44	Long-term financial facilities received	x113	Gross profit on sales or revenue

Table 3: continue

x52	Equity of long-term assets	x121	Working capital to fixed assets
x53	Accounts paid on sales revenue	x122	Fixed assets to total assets
x54	Cash flow sold or earned	x123	Receipts to total debts
x57	Cash flow to total debt	x124	Financial cost to receive
x58	Long-term amount to total assets	x125	Financial costs to current debt
x61	Short-term debt to total expenses	x130	Logarithm of total assets
x62	Current debt to net sales	x131	Debt to banks
x63	Current liabilities to total assets	x132	Inventory of materials and goods * 365 sold or earned
x69	Current assets - Inventory Long-term assets}	x138	Inventories

2.2 Logistic Regression Model

The correlation coefficient is often used to express the intensity of the linear relationship between two quantitative variables. The regression model is also employed to display the relationship model between the two. In the meantime, a model is created to predict the dependent variable (Y) based on the independent variable (X) [31]. However, it should be noted that both independent and dependent variables in the created model are quantitative. In addition, the condition of correlation of these values lies in the regression method. However, scholars may be interested in measuring the relationship between an independent variable (with continuous values) and a dependent variable with qualitative values. In this case, the normal linear regression method will not work out and it is compulsory to use the "logistic regression" method. The logistic regression model can be seen as a specific case of the general linear and linear regression model. The logistic regression model is based on quite different hypotheses (about the relationship between dependent and independent variables) of the linear regression. An important difference between these two models can be attributed to the two features of logistic regression. First, the conditional distribution $y|\vec{x}$ is a Bernoulli distribution instead of a Gaussian distribution since the dependent variable is a binary one. Second, the prediction values are probabilistic and limited to the range between zero and one and they can be obtained using the logistic distribution function. The logistic regression predicts the odds of an outcome. It is known that linear regression refers to creating a parametric linear relationship to represent the relationship between an independent variable and a dependent variable. The form of a simple linear regression model is as follows:

$$Y=\beta_0+\beta_1 X+\epsilon \quad (1)$$

As can be seen, this relation is the equation of a line to which the error sentence or "i" is added. The parameters of this linear model are the "y-intercept" (β_0) and the gradient of the line (β_1). In this case, if "$\hat{Y}$" is the estimated value for the dependent variable, it can be considered as the mean of observations for the dependent variable for the constant value of the independent variable. So if we replace the mean with the expected value, assuming that the mean of the error sentence is also zero, we will have:

$$\hat{Y} = E(Y|X = x) = \hat{\beta}_0 + \hat{\beta}_1 x \quad (2)$$

where E(Y|X=x) indicates the conditional (average) expected value, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are also the estimates related to each of the parameters. If the value of the dependent variable (Y) is binary (two states) and contains 0 and 1, it clearly has a Bernoulli distribution and its mathematical expectation is calculated as follows:

$$\hat{Y} = E(Y|X = x) = P(Y = 1|X = x) = p(x) \quad (3)$$

Hence, the regression model for the Bernoulli dependent variable is determined. The prediction value for the dependent variable was made with the probability of p(x). We needed a function that varies from about 0 to 1 to determine the relationship model between the dependent and the independent variable instead of a linear one. In logistic regression, where the dependent variable is two-dimensional, the effect of independent variables on the dependent variable is shown as the role of each independent variable in the probability of occurrence of a particular class of the dependent variable. In logistic regression, the probability of occurring a particular class of the dependent variable, which is called the probability of an event, is estimated based on the exponential function of independent variables. The domain of this function is real numbers and its range is between zero and one. The function is introduced below and its plotted related diagram based on the parameters b_0=0 and b_1=1 can be seen in Figure 1.

$$f(x) = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} \tag{4}$$

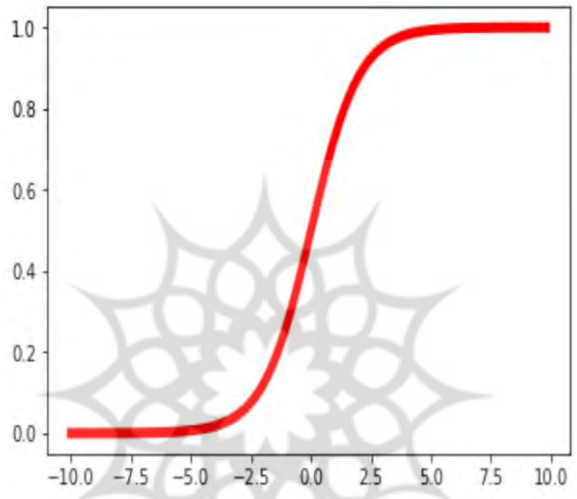

Fig. 1: The standard logistic function

2.3 Support Vector Machine (SVM)

The support vector machine (SVM) can be considered a supervised learning approach in which input-output mapping functions are produced from a set of labeled training data. The mapping function can be either a regression function or a classification function [32]. The classification function is actually the category of the input data. Nonlinear kernel functions are frequently utilized for the case of classification to convert input data into a high-dimensional feature space. Under such a circumstance, in comparison to the original input space, the input data become more separable. Then, maximum-margin hyperplanes are engendered. Therefore, the proposed model is dependent on just a subset of the training data close to the class boundaries. Similarly, any training data that is sufficiently close to the model prediction is ignored by the produced model by the support vector regression. Besides, SVMs are told to belong to "kernel methods". The support vector machine has shown greatly competitive performance in many real-world applications like image processing, face recognition, text mining, and bioinformatics in addition to its solid mathematical foundation in statistical learning theory. This has introduced SVMs as one of the state-of-the-art tools for data mining and machine learning, along with other soft computing approaches, e.g., fuzzy systems and neural networks. The main purpose of the SVM algorithm is the creation of the best line or decision boundary that is able to segregate n-dimensional space

into classes. Under such a circumstance, the new data point is easily put in the correct category in the future. This best decision boundary is called a hyperplane. SVM selects the extreme vectors/points that help in engendering the hyperplane. These extreme cases are called support vectors, and thus, the algorithm is named SVM. Consider the below diagram in which two different categories are classified using a decision boundary or hyperplane:

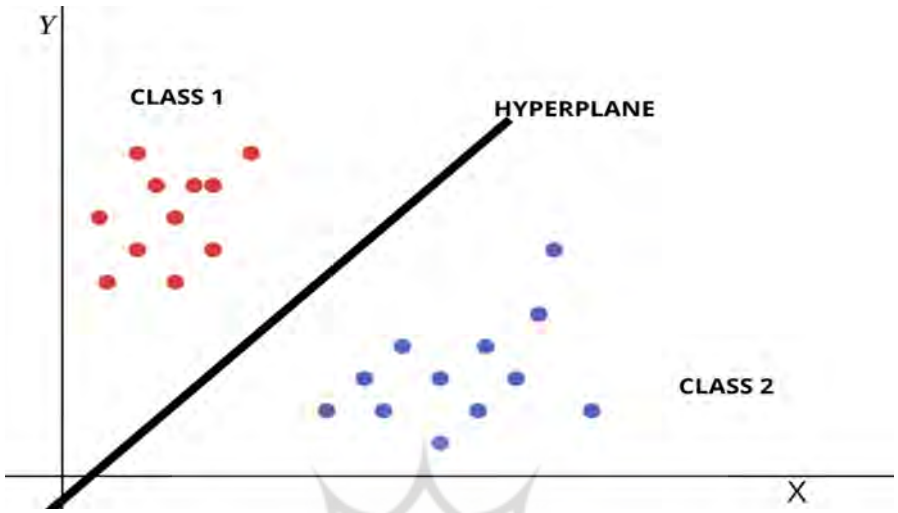

Fig. 2: The Support Vector Machines (SVM) Method

2.4 Gradient Boosting (XGBoost) Model

The gradient boosting method uses the ‘weak’’ decision trees similar to the Random Forest. The main difference between the two methods is that the trees are trained one after the other in the gradient boosting method [33]. Each subset tree is primarily trained with the data that has been mistakenly predicted by the previous tree. This makes the model less focused on issues that are easy to predict and more focused on complex ones. Gradient boosting is a machine learning method for regression and classification problems, which creates a predictive model in the form of a set of weak predictive models. The XGBoost method was also used in this research. Like other boosting methods, gradient boosting is a linear combination of a series of weak models to develop a strong and efficient model.

2.5 Ensemble Stacking Method

Ensemble machine learning models or ensemble models are one of the machine learning approaches, in which, several models called weak learner or base models, are trained to solve a problem and are combined aimed at achieving better results. The combined stacking method combines base models using the meta-model, while the bagging method uses deterministic algorithms to combine base models with each other. [34] As mentioned earlier, the main idea of stacking is to train several different base models and combine them through training a meta-model to perform the final prediction based on the forecasts made by the base models. Thus, building a stacking model requires two things: L basic models to train the data and a meta-model for combining the results. Figure 3 illustrates the outline of the stacking method:

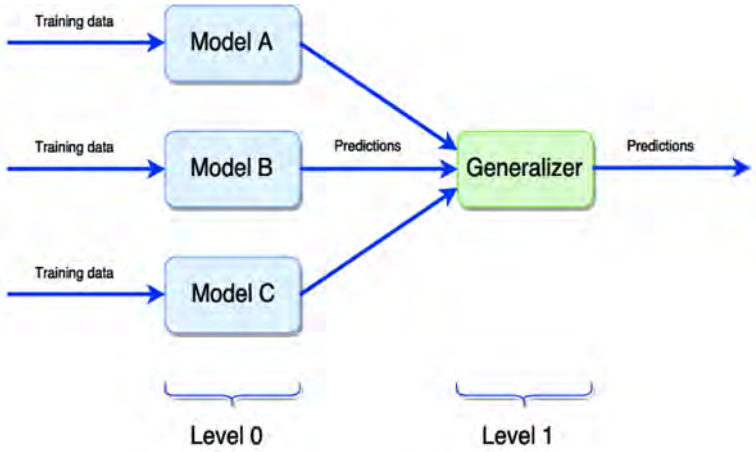

Fig. 3: The Ensemble Stacking Method

As mentioned, we divided the database into two parts since the predictions made on the data used to train the base models should not be used for the meta-model training. In the end, an ensemble stacking model was trained using 3 base models and the optimized features.

3 Results and Discussion

3 base models (logistic regression, SVM, and XGBoost) were used in this research for modeling. Finally, the ensemble stacking model was used to train the models. To do so, the data was divided into two parts: training and testing. The training data were financial statements before 2016, including 9273 data. The test data were related to financial statements after 2016, which included 3765 data records. The obtained results for each method are documented in Tables 4 to 11.

≠ Results with SVM

Table 4: Evaluation of test data using the SVM model

CLASS	PRECISION	RECALL
0	0.77	0.76
1	0.86	0.82
AVG/TOTAL	0.81	0.81

Table 5: Confusion Matrix

Actual/Prediction	0	1
0	1579	285
1	520	1381

AUC: 0. 7396

Accuracy Ratio :0.7897

≠ Results with logistic regression

Table 6. Evaluation of Test Data Using the Lr Model

CLASS	PRECISION	RECALL
0	0.81	0.86
1	0.87	0.79
AVG/TOTAL	0.83	0.83

Table 7: Confusion Matrix

Actual/Prediction	0	1
0	1560	300
1	439	1466

AUC: 0.8097

Accuracy Ratio: 0.6837

≠ **Results with XGBoost**

Table 8: Evaluation of test data using XGBoost model

CLASS	PRECISION	RECALL
0	0.74	0.76
1	0.81	0.83
AVG/TOTAL	0.78	0.78

Table 9: Confusion Matrix

Actual/Prediction	0	1
0	1557	307
1	487	1414

AUC: 0.8890

Accuracy Ratio: 0.7267

≠ **Results with Ensemble learning**

The evaluation of the ensemble learning (stacking) model on the test data using the features selected by each of the basic models is as follows.

Table 10. Evaluation of Test Data Using Stacking Model

Class	Precision	Recall
0	0.86	0.87
1	0.85	0.86
AVG/TOTAL	0.86	0.86

Table 11: Confusion Matrix

Actual/Prediction	0	1
0	1565	299
1	289	1612

AUC: 0.9276

Accuracy Ratio: 0.8247

As it was observed in this research, the combined stacking method has achieved a much better result. The stacking technique is a kind of improved version of the voting technique. In this technique, each model has a different share in the decision-making process and this share in the training process is determined by the decision-making model. The prediction accuracy of a ensemble learning model compared to other machine learning models depends on several factors, such as the size and complexity of the dataset, the quality of the data, and the hyperparameters of the models. However, in general, collective learning models have been shown to improve prediction accuracy when compared to single models in several studies. One reason for the improved accuracy of collective learning models is that they can

reduce overfitting, which is a common problem in machine learning. The ensemble of models is able to generalize better by reducing the variance of the predictions. Additionally, collective learning models can capture different aspects of the data, which can improve the overall accuracy of the predictions. Overall, collective learning models have been shown to be effective at improving prediction accuracy compared to single models in several studies. However, the specific performance of these models compared to other machine learning models depends on the specific characteristics of the dataset and the task at hand. The comparison of previous researches on corporate bankruptcy prediction based on accuracy is presented in Table 12.

Table 12: Comparison of previous researches on corporate bankruptcy prediction

Reference	**Year**	**Data set**			**Method**	**Accuracy (%)**
		Country	number of corporate	Features		
[35]	2019	Greece	100	20	Deep neural networks	73
[36]	2019	French	2150	50	Boosting	75
[37]	2018	Polish	1000	15	Jordan recurrent neural	81
[38]	2018	Polish	10503	64	ANN	88
[39]	2018	Colombia	1000	15	Decision tree	91
[40]	2017	American and Canadian	998	11	Random forest	87
[41]	2016	Taiwan	480	190	KNN	82

4 Conclusions

One of the main objectives of the current research paper is to utilize ensemble methods in bankruptcy prediction instead of just using a classification method. In the field of finance, bankruptcy forecasting can be considered one of the most momentous research because the possibility of correct bankruptcy can be used and done in a location that can reduce the high costs of bankruptcy. Reputable rating agencies and banks typically utilize this model to assign ratings and make credit decisions. Predicting bankruptcy and subsequently rooting out the problem and having a solution for that can engender very satisfying results. Looking at various types of financial processes and methods that concentrate on historical information is one way to perform financial analysis. Due to the fact that the financial ratios are from the loss and profit statement of the business unit and the items in the balance sheet, therefore, some financial ratios from the unification of the historical financial statements are significantly different from the historical items. The accuracy rate of models can be considerably improved as confirmed by research findings. Machine learning techniques are shown remarkably advanced prediction ability in different scientific fields like the financial sector. Identification of the pertinent characteristics that drives this prediction and improves explanation ability is a challenging task in this domain. In the current research, an attempt was made to design the bankruptcy forecasting model by employing all the financial ratios utilized in various models in financial bankruptcy forecasting. Thus, according to the prediction methods used in this research, the WOE method is used to identify the most important prediction variables. The lowest error rate was given by the results obtained from the confusion matrix with a similar subset of any vote. This demonstrates that the single approach to feature election can restrict machine learning algorithms in achieving optimal decision-making process and also can enhance the error level in the study. In addition, it can be seen that the XGboost algorithm achieves the highest AUC scores among the three used basic predictor models which makes it a cost-effective and efficient algorithm. Besides, it was shown that the ensemble approach to bankruptcy forecasting outperformed the single learning. This can be considered supportive proof for the available hypothesis that

ensemble methods produce much higher performance in comparison to single learners. In the general form, all the experiments have given a great performance compared to existing studies in the open literature in this domain. The results confirm the high accuracy of the Stacking hybrid model in predicting the financial bankruptcy risk of listed and OTC corporations. The stacking technique, which is one of the most powerful techniques of collective learning, shows great success in such issues. Regarding the high power of the produced model and also stated research findings in this study, it can be concluded that the model can be successfully used by investors in choosing the optimal portfolio and as creditors to help prevent lending to corporations with high bankruptcy risk. Generally speaking, it can be said that by using the results of this study as a first step and also performing preventive actions, corporations can be prevented from bankruptcy and future losses can be reduced.

Reference

[1] H. Reinert, E.S. Reinert, *Creative destruction in economics: Nietzsche, Sombart, Schumpeter*, in: Friedrich Nietzsche (1844–1900), Springer, 2006; 55-85. doi: 10.1007/978-0-387-32980-2_4

[2] N. Chen, B. Ribeiro, A.S. Vieira, J. Duarte, J.C. Neves, A genetic algorithm-based approach to cost-sensitive bankruptcy prediction, *Expert Systems with Applications*, 2011; 38: 12939-12945. doi: 10.1016/j.eswa.2011.04.090

[3] Surwanti, A., Ramdan F., and Rosnia M., Predicting Corporate Bankruptcy in Indonesia's Transportation Industry, *Journal Aplikasi Manajemen*, 2022; 20(2):276-288.

[4] Alanis, E., Sudheer Ch., and Agam, Sh., Benchmarking machine learning models to predict corporate bankruptcy, *arXiv preprint arXiv:2212.12051*, 2022. doi:10.48550/arXiv.2212.12051

[5] Lombardo, G., et al., Machine Learning for Bankruptcy Prediction in the American Stock Market: Dataset and Benchmarks, *Future Internet* 2022; 14(8): 244.doi: 10.3390/fi14080244

[6] Lahmiri, S., Stelios, B., Can machine learning approaches predict corporate bankruptcy? *Evidence from a qualitative experimental design*, *Quantitative Finance*, 2019; 19(9): 1569-1577. doi: 10.1080/14697688.2019.1588468

[7] Kim, H., Hoon, C., and Doojin, R., Corporate default predictions using machine learning: Literature review, *Sustainability*, 2020; 12(1): 66325. doi: 10.3390/su12166325

[8] Meese, E.N., Viken, T., Machine Learning in Bankruptcy Prediction, *Norwegian School of Economics Bergen*, Spring 2019.

[9] Beaver, W.H., Financial ratios as predictors of failure, *Journal of accounting research*, 1966; 71-111. doi: 10.2307/2490171

[10] Beaver, W.H., Alternative accounting measures as predictors of failure, *The accounting review*, 1968; 43:113-122.

[11] Beaver, W.H., Kennelly, J.W., Voss, W.M., Predictive ability as a criterion for the evaluation of accounting data, *The Accounting Review*, 1968; 43: 675-683. doi: 10.1111/j.1467-6281.1971.tb00390.x

[12] Boyacioglu, M.A., Kara, Y. Baykan, Ö.K., Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit

insurance fund (SDIF) transferred banks in Turkey, *Expert Systems with Applications*, 2009; 36: 3355-3366.

[13] Cetina, J., Loudis, B., The influence of systemic importance indicators on banks' credit default swap spreads, *Journal of Risk Management in Financial Institutions*, 2016; 9: 17-31.

[14] Cai, L., Singenellore, R., Bloomberg Credit Default Risk for Private Corporations, *Bloomberg Professional Service*, 2015.

[15] R.C. Merton, On the pricing of corporate debt: The risk structure of interest rates, *The Journal of finance*, 1974; 29: 449-470.

[16] Zhang, B.Y., Zhou H., Zhu, H., Explaining credit default swap spreads with the equity volatility and jump risks of individual firms, *The Review of Financial Studies*, 2009; 22: 5099-5131.

[17] Crouhy, M., Galai, D., Mark, R., A comparative analysis of current credit risk models, *Journal of Banking & Finance*, 2000; 24: 59-117. doi: 10.1016/S0378-4266(99)00053-9

[18] Næss, A.B., Wahlstrøm, R.R., Helland, F.F., Kjærland, F., Konkursprediksjon for norske selskaper-En sammenligning av regresjonsmodeller og maskinlæringsteknikker, Bred og spiss! *NTNU Handelshøyskolen 50 år: En vitenskapelig jubileumsantologi*, 2017.

[19] Wahlstrøm, R.R., Helland, F.F., Konkursprediksjon for norske selskaper–en analyse ved maskinlæringsteknikker og tradisjonelle statistiske metoder, in, 2016.

[20] Bernhardsen, E., A model of bankruptcy prediction, *Working Paper*, 2001.

[21] Berg, D., Bankruptcy prediction by generalized additive models, *Applied Stochastic Models in Business and Industry*, 2007; 23: 129-143.

[22] Bae, J.K., Predicting financial distress of the South Korean manufacturing industries, *Expert Systems with Applications*, 2012; 39: 9159-9165. doi: 10.1016/j.eswa.2012.02.058

[23] Altman, E.I., Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The journal of finance*, 1968; 23: 589-609. doi: 10.2307/2978933

[24] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Least angle regression, *The Annals of statistics*, 2004; 32: 407-499.doi: 10.1214/009053604000000067

[25] Komijanie, A., Sa'adatfar, J., Application of artificial neural network models in predicting economic bankruptcy of the registered corporations in stock market, *Journal of Iran's Economic Essays*, 2006; 3:1-45.

[26] Philosophov, L.V., Corporate bankruptcy prognosis: An attempt at a combined prediction of the bankruptcy event and time interval of its occurrence, *International Review of Financial Analysis*, 2002; 11: 375-406. Doi: 10.1016/S1057-5219(02)00081-9

[27] Tone, K., Toloo, M., Izadikhah, M., A modified slacks-based measure of efficiency in data envelopment analysis, *European Journal of Operational Research*, 2020; 287:560-571. doi: 10.1016/j.ejor.2020.04.019

[28] Xie, C., Luo, C., Yu, X., Financial distress prediction based on SVM and MDA methods: the case of Chinese listed corporations, *Quality & Quantity*, 2011; 45: 671-686.

[29] Lee, M.-C., To, C., Comparison of support vector machine and back propagation neural network in evaluating the enterprise financial distress, *arXiv preprint arXiv*:1007.5133, 2010. doi:10.48550/arXiv.1007.5133

[30] Dibachi, H., Behzadi, M., Izadikhah, M., Stochastic modified MAJ model for measuring the efficiency and ranking of DMUs, *Indian Journal of Science and Technology*, 2015; 8: 1-7.doi: 10.17485/ijst/2015/v8iS8/71505

[31] Tolles, J., Meurer, W.J., Logistic regression: relating patient characteristics to outcomes, *Jama*, 2016; 316: 533-534. doi: 10.1001/jama.2016.7653

[32] Ben-Hur, A., Weston J., A user's guide to support vector machines, in: *Data mining techniques for the life sciences, Springer*, 2010; 223-239. doi: 10.1007/978-1-60327-241-4_13

[33] Sagi, O., Rokach, L. Approximating XGBoost with an interpretable decision tree, *Information Sciences*, 2021; 572: 522-542. doi: 10.1016/j.ins.2021.05.055

[34] Alfaro-Navarro, J.-L., Cano, E.L., Alfaro-Cortés, E., García, N., Gámez, M., Larraz, B., A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems, *Complexity*, 2020. doi: 10.1155/2020/5287263

[35] Alexandropoulos, S.-A.N., Aridas, C.K., Kotsiantis, S.B., and Vrahatis, M.N., A Deep Dense Neural Network for Bankruptcy Prediction. *International Conference on Engineering Applications of Neural Networks*. Springer. 2019.doi: 10.1007/978-3-030-20257-6_37

[36] du Jardin, P., Veganzones, D., and Séverin, E., Forecasting corporate bankruptcy using accrual-based models. Comput. Econ, 2019; 54: 7–43. doi: 10.1007/s10614-017-9681-9

[37] Hardinata, L., Warsito, B., Bankruptcy Prediction Based on Financial Ratios Using Jordan Recurrent Neural Networks: A Case Study in Polish Companies, *Journal of Physics: Conference Series.* IOP Publishing, 2018.

[38] pozorska, J., and Scherer, M., Company Bankruptcy Prediction with Neural Networks. *International Conference on Artificial Intelligence and Soft Computing.* Springer 2018.doi: 10.1007/978-3-319-91253-0_18

[39] Arroyave, J., A comparative analysis of the effectiveness of corporate bankruptcy prediction models based on financial ratios: *Evidence from Colombia. J. Int. Studies*., 2018; 11: 273–287.

[40] Barboza, F., Kimura, H., and Altman, E., Machine learning models and bankruptcy prediction, *Expert Syst. Appl*., 2017; 83: 405–417. doi: 10.1016/j.ejor.2016.01.012

[41] iang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A., Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *Eur. J. Oper. Res*., 2016; 252:561–572. doi: 10.1016/j.ejor.2016.01.012