


Article

Comparing the Performance of Corporate Bankruptcy Prediction Models Based on Imbalanced Financial Data

Seol-Hyun Noh 

Department of Statistical Data Science, ICT Convergence Engineering, Anyang University,
Anyang 14028, Republic of Korea; shnoh@anyang.ac.kr

Abstract: Forecasts of corporate defaults are used in various fields across the economy. Several recent studies attempt to forecast corporate bankruptcy using various machine learning techniques. We collected financial information on 13 variables of 1020 companies listed on the KOSPI and KOSDAQ to capture the possibility of corporate bankruptcy. We propose a data processing method for small-sample domestic corporate financial data. We investigate the case of random sampling of non-bankrupt companies versus sampling non-bankrupt companies based on approximate entropy and optimized threshold based on AUC to address the imbalance between the number of bankrupt companies and the number of non-bankrupt companies. We compare the performance measures of corporate bankruptcy prediction models for the small sample data structured in two ways and the full dataset. The experimental results of this study contribute to the selection of an appropriate corporate bankruptcy prediction model.

Keywords: corporate bankruptcy; bankruptcy prediction; performance comparison; imbalanced financial data



Citation: Noh, S.-H. Comparing the Performance of Corporate Bankruptcy Prediction Models Based on Imbalanced Financial Data. *Sustainability* **2023**, *15*, 4794. <https://doi.org/10.3390/su15064794>

Academic Editors: Albert Y.S. Lam and Yanhui Geng

Received: 2 February 2023

Revised: 4 March 2023

Accepted: 6 March 2023

Published: 8 March 2023



Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Many sectors of the economy use predictions of corporate bankruptcy. A bankruptcy prediction model allows companies to diagnose the current state and establish strategies and management to operate the business more stably by managing key indicators that affect its bankruptcy risk. It also allows investors to establish investment strategies and improve their portfolios and governments to establish policies to improve macroeconomic soundness and financial regulatory policies.

Several recent studies use various methods to predict corporate bankruptcy. In particular, since the 1990s, computational economics focused on predicting corporate bankruptcy with artificial neural networks [1–10].

Barboza et al. [3] compare machine learning models with traditional models for bankruptcy prediction and show that machine learning models had 10% better performance in accuracy on average. Specific machine learning bankruptcy prediction models include classification models such as support vector machines (SVM), logistic regression (LR), k-nearest neighbors (k-NN), decision tree (DT), and random forest (RF), and time series models such as recurrent neural network (RNN) and long short-term memory (LSTM). These models are widely used as consumer bankruptcy prediction models and corporate bankruptcy prediction models [11–14].

Classification is a supervised learning technique to conduct predictive analytics with a categorical outcome, whether binary or multiclass. Much current research concentrates on classification using several algorithms from basic to advanced, such as logistic regression, discriminant analysis, Naïve Bayes, decision tree, random forest, support vector machine, neural network, and so on [15]. These are well developed and successfully applied to many application domains. However, the majority of supervised learning techniques were developed for balanced class distribution, leaving imbalanced class distribution relatively

neglected. For this reason, many researchers seek to address the imbalance between the number of bankrupt companies and the number of non-bankrupt companies.

Kim et al. [7] use RNN and LSTM models to predict bankruptcy and compare their performance with other classification models, including of SVM, LR, and RF. From experiments, Kim et al. [7] demonstrate that the RNN and LSTM models outperformed in terms of accuracy, precision, recall, and AUC, using the monthly financials from the Compustat North America Dataset for about 45,472 non-financial firms for January 2007 to December 2019. Since 2057 of the firms went bankrupt, the authors had a large amount of monthly financial information on the bankrupt companies and thus had a sufficient training dataset for the financial information of the bankrupt companies for time series models such as RNN and LSTM. As a balanced classification, accuracy may be the unbiased metric for evaluation. Therefore, a corporate bankruptcy prediction model suitable for unbalanced corporate financial data is necessary.

In this study, we propose a random small sampling method of non-bankrupt companies and a method of small sampling of companies representing non-bankrupt companies based on approximate entropy to improve corporate bankruptcy prediction performance for imbalanced corporate financial data. To deal with the imbalanced class, we also propose an optimal threshold method to improve AUC performance even when considering the imbalanced total dataset. The optimal threshold method proposed in this study applies the concept of the anomaly score defined in [16] to corporate bankruptcy prediction models. We provide experimental results to analyze the effectiveness of models for predicting corporate bankruptcy for imbalanced corporate financial information in Section 3.

To derive these results, we established the following research questions:

- Research question 1: How can we derive data sampling methods that improve the performance of corporate bankruptcy prediction models for imbalanced corporate financial information?
- Research question 2: How can we derive an optimal threshold technique that improve AUC performance even when considering the imbalanced corporate financial information?

In order to solve the above two research questions, the research results were derived according to the research process as shown in Figure 1.

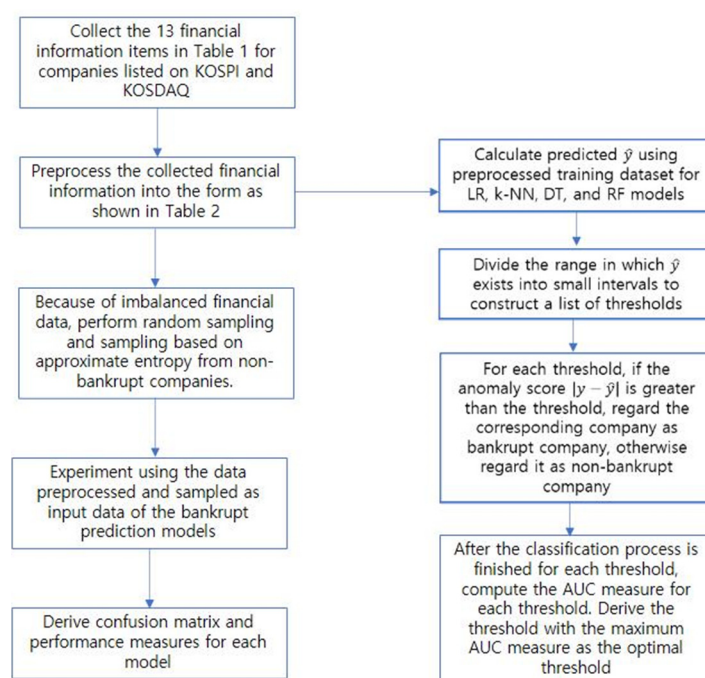


Figure 1. Diagram of research stage.

2. Data and Methods

2.1. Data and Sampling

2.1.1. Data

We collected the 13 financial information items for about 1020 companies listed on the Korea Composite Stock Price Index (KOSPI) and Korean Securities Dealers Automated Quotations (KOSDAQ) from 2012 to 2021. As 20 companies went bankrupt, it is not possible to build a sufficient training dataset for the RNN and LSTM time series models. Table 1 summarizes the items collected for this sample, excluding the parent company equity holder item. The corporate financial information in Table 1 corresponds to the criteria that characterize bankrupt and non-bankrupt companies [2]. Cha and Kang [2] conducted a multivariate discriminant analysis to select the 29 variables like in Appendix A that correspond to the criteria characterizing bankrupt and non-bankrupt companies from a set of continuous independent variables of financial information. After calculating the correlation coefficient of these 29 variables to create highly correlated variable groups, we conducted a t-test to detect statistically significant differences between the population mean of bankrupt and non-bankrupt companies within each variable group. Table 1 presents the resulting 14 variables with the lowest p-values. We collected the 13 financial information items from FnGuide through the FnDB Navigator.

Table 1. Financial information collected for each sample company (excluding the parent company equity holder item. Source: [2]).

Financial Information	<i>t</i> -Test <i>p</i> -Value
Total assets	2.33×10^{-5}
Parent company equity holder	3.59×10^{-4}
Intangible assets ratio	4.84×10^{-2}
Equity capital ratio	1.48×10^{-20}
Debt ratio	9.63×10^{-6}
Cash flows/total liabilities	2.23×10^{-8}
Total assets growth rate	2.88×10^{-14}
Operating revenue/operating expense	6.71×10^{-13}
Gross margin	2.45×10^{-4}
ROA (income from continuing operations before tax)	1.79×10^{-9}
ROA (operating profit)	2.06×10^{-10}
ROE (income from continuing operations before tax)	2.89×10^{-5}
ROE (operating profit)	2.52×10^{-2}
Total debt turnover ratio	5.68×10^{-27}

Among the 1020 listed companies, we define those delisted between 2012 and 2021 as bankrupt and used the financial data of the 20 bankrupt companies and 1000 non-bankrupt companies as the input for the LSTM, LR, k-NN, DT, and RF bankruptcy prediction models. To test bankruptcy prediction performance while preventing overfitting in the models caused by the data of non-bankrupt companies due to the small number of bankrupt companies, we conducted an experiment using the data of all companies, a subsample of 10 bankrupt companies, and a subsample of 20 non-bankrupt companies for the training dataset and test dataset. Table 2 summarizes the collected financial information configured as the input and label data of the model. Figure 2 shows the input data and label data of LR, k-NN, DT, and RF bankruptcy prediction models. For companies that went bankrupt in 2013, financial data for 2013 were padded and used as financial data from 2014 to 2021, and for companies that went bankrupt in 2020, financial data for 2020 were used as financial data for 2021. We applied the same padding technique to label data [17].

Table 2. Input and label data in the proposed bankruptcy prediction model.

LSTM input data: Company (Financial information in 2012, . . . , 2020)
LSTM label data: Company (Bankruptcy or non-bankruptcy): (1)/(0)
LR, k-NN, DT, RF input data: Company (Financial information in 2012) ⋮ (Financial information in 2020)
LR, k-NN, DT, RF label data: Company (Bankruptcy or non-bankruptcy in 2013) ⋮ (Bankruptcy or non-bankruptcy in 2021)

node_id	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	time
924	49.73	64.10	−5.43	18.37	29.08	110.42	8.32	−6.08	16.32	−11.94	1.80	38,179,277	6,062,267	2012
984	67.94	18.59	−15.54	23.56	19.92	93.37	−3.04	−7.07	−3.94	−9.14	1.89	25,364,197	3,365,951	2012
⋮														⋮
511	39.17	108.57	15.10	96.21	30.45	106.42	7.99	3.48	18.53	8.06	2.33	16,634,320	526,704	2012
512	73.36	12.03	14.68	10.94	5.96	99.40	−0.90	2.22	−1.18	2.89	6.49	149,560,052	1895	2012
⋮														⋮

(a) Input data of LR, k-NN, DT, and RF bankruptcy prediction models

node_id	2013	2014	2015	2016	2017	2018	2019	2020	2021
924	0	0	0	1	1	1	1	1	1
984	0	1	1	1	1	1	1	1	1
⋮									
511	0	0	0	0	0	0	0	0	0
512	0	0	0	0	0	0	0	0	0
⋮									

(b) Label data of LR, k-NN, DT, and RF bankruptcy prediction models**Figure 2.** Input data and label data of the models. In panel (a), v1 to v13 refer to the 13 financial information items.

As we must configure the input data for the LSTM model with time series data in an experiment on all of the 1020 companies, the training and testing datasets were configured with data from 10 bankrupt companies and 500 non-bankrupt companies for the experiment. For the LR, k-NN, DT, and RF models, we set the financial information in the n -th year as the input data and bankruptcy or non-bankruptcy in year $n + 1$ as the label data, creating 9 times more input data than the LSTM model for each company. Hence, the training and testing datasets were configured with the data of 90 bankrupt companies and 4500 non-bankrupt companies for the experiment. As the input data for the LR, k-NN, DT, and RF models have 9 times as many bankrupt companies as the LSTM model, we expect better bankruptcy prediction performance.

With data from a small sample of bankrupt companies, overfitting occurs for non-bankrupt companies in the LR, k-NN, DT, and RF models when conducting an experiment using all companies. As a result, bankruptcy prediction performance of the models decreases. Therefore, considering the number of bankrupt companies, we set the financial information of 10 bankrupt companies and 20 non-bankrupt companies as small sample data consisting of the training and testing datasets and conducted a comparison with an experiment using all 1020 companies. For a small sample, we randomly sampled the non-bankrupt companies 5 times and report the mean performance measure of the 5 repeated experiments.

Similarly, we set the training and testing datasets for the LSTM model using the data of 10 bankrupt companies and 20 non-bankrupt companies for the experiment.

2.1.2. Sampling

In the small-sample data experiment, we used the financial information of bankrupt 10 companies and 20 non-bankrupt companies as the training dataset. We found performance differences depending on the sampling method used to select the 20 non-bankrupt companies from the 1000 non-bankrupt companies, as described in Section 3. To reduce this difference, we conducted random sampling of the non-bankrupt companies 5 times and used the mean performance measure of 5 repeated experiments. As data imbalances may still occur using this method, we set the time series for each of the 13 financial information items and compared the performance with the sampling of 20 non-bankrupt companies which we consider as having characteristics that represent non-bankrupt companies as their entropy is low.

The approximate entropy we used is a method that measures the entropy of a time series, with the following definition [18]: we configure $N - m + 1$ vectors $x_m(i) = [u(i + k) : 0 \leq k \leq m - 1]$ for time series $[u(j) : 1 \leq j \leq N]$. The number of vectors $x_m(j)$ within the distance of r from $x_m(i)$ is set as B_i and the number of vectors $x_{m+1}(j)$ within the distance of r from $x_{m+1}(i)$ as A_i . m is the length of compared sequences, and r is a matching tolerance, where the constant is determined before measuring approximate entropy. The distance between two vectors $x_m(i)$ and $x_m(j)$ is

$$d(x_m(i), x_m(j)) = \max\{|u(i + k) - u(j + k)| : 0 \leq k \leq m - 1\},$$

where $C_i^m(r) = \frac{B_i}{N - m + 1}$, $C_i^m(r)$ is the probability that vector $x_m(j)$ exists within the distance of r with $x_m(i)$.

Approximate entropy $ApEn(m, r, N)$ is

$$\frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \ln|C_i^m(r)| - \frac{1}{N - m} \sum_{i=1}^{N-m} \ln|C_i^{m+1}(r)|,$$

which is approximately $\frac{1}{N-m} \sum_{i=1}^{N-m} \left\{ -\ln\left(\frac{A_i}{B_i}\right) \right\}$.

Companies with a low approximate entropy across financial information items are considered to represent the characteristics of non-bankrupt companies, as they have a low degree of disorder across items. We sampled 20 non-bankrupt companies in the training dataset in descending order of approximate entropy and compared the results with an experiment using the data on a random, small sample. The mean approximate entropy was measured for the time series of each financial information item, and the approximate entropy of 20 sampled non-bankrupt companies was within 1.70. To measure approximate entropy $m = 3$, we set r as the standard deviation of each financial information item.

2.2. Models and Performance Measures

We tested the LSTM, LR, k-NN, DT, and RF models as corporate bankruptcy prediction models. The LR, k-NN, DT, and RF models used the Keras Library, and the LSTM model was implemented using Python 3.8 for Ubuntu Linux as follows.

2.2.1. LSTM Model

The LSTM model is a deep learning model that uses time-series data as the input and outputs future data, and it is used to predict river levels, solar power generation, fine dust, energy demand, and stock prices [19]. Using the dataset in Section 2.1.1 as the input and label data, we establish a model to predict bankruptcy in the next year. The mathematical model of the LSTM is expressed as in Equation (1) and Figure 3. Output h_t , outputgate o_t , memory cell c_t , new memory content \tilde{c}_t , forget gate f_t , and input gate i_t in the LSTM are modeled as in Equation (1) and illustrated in Figure 2. The sigmoid function $\sigma(x)$ is

defined as $\sigma(x) = \frac{1}{1+e^{-x}}$ for real value x . If X is a matrix with real numbers as elements, $\sigma(X)$ is a matrix with the values of the sigmoid function for each element of X as elements

LSTM model pseudocode

```

Step (1)
train_corp_data = 9-year data of 13 financial information items of 10 bankrupt companies
test_corp_data = 9-year data of 13 financial information items of the other 10 bankrupt companies
train_corp_label = whether a company went bankrupt during 6 years, obtained by setting 4 years
as the sequence length for train_corp_data
test_corp_label = whether a company went bankrupt during 6 years, obtained by setting 4 years
as the sequence length for test_corp_data
Step (2)
Normalizing 4 datasets in Step 1
Step (3)
Stacking 4 years of financial information corresponding to the sequence length as input data for
each LSTM
Step (4)
Setting the hidden size of the LSTM at 256 and building the LSTM model using nn.LSTM()
provided from torch.nn
Step (5)
Store the output value of the LSTM in the output variable, while forwarding the input data of the
LSTM through the device.
Step (6)
Calculating a performance measure by outputting the confusion matrix

```

$$\text{Output } h_t = o_t \tanh(c_t) \quad (1)$$

$$\text{Output gate } o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t)$$

$$\text{Memory cell } c_t = f_t c_{t-1} + i_t \tilde{c}_t$$

$$\text{New memory content } \tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1})$$

$$\text{Forget gate } f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1})$$

$$\text{Input gate } i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1})$$

2.2.2. Logistic Regression Model

The logistic regression model is a method of modeling the conditional probability $P(Y = 1|x_1, x_2, \dots, x_k)$ as shown in Equation (2), where X_1, X_2, \dots, X_k are explanatory variables and Y (Y is 0 or 1) is the binary response variable.

$$\log \left[\frac{P(Y = 1|x_1, x_2, \dots, x_k)}{1 - P(Y = 1|x_1, x_2, \dots, x_k)} \right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (2)$$

From the training dataset, intercept α and the effect of x_i , namely β_i ($i = 1, 2, \dots, k$), are estimated with maximum likelihood estimation.

It is an algorithm that solves the problem of binary classification as it considers new data as $Y = 1$ if $P(Y = 1|x_1, x_2, \dots, x_k)$ obtained from Prediction Equation (2) is greater than the pre-defined threshold and $Y = 0$ if it is smaller than the threshold.

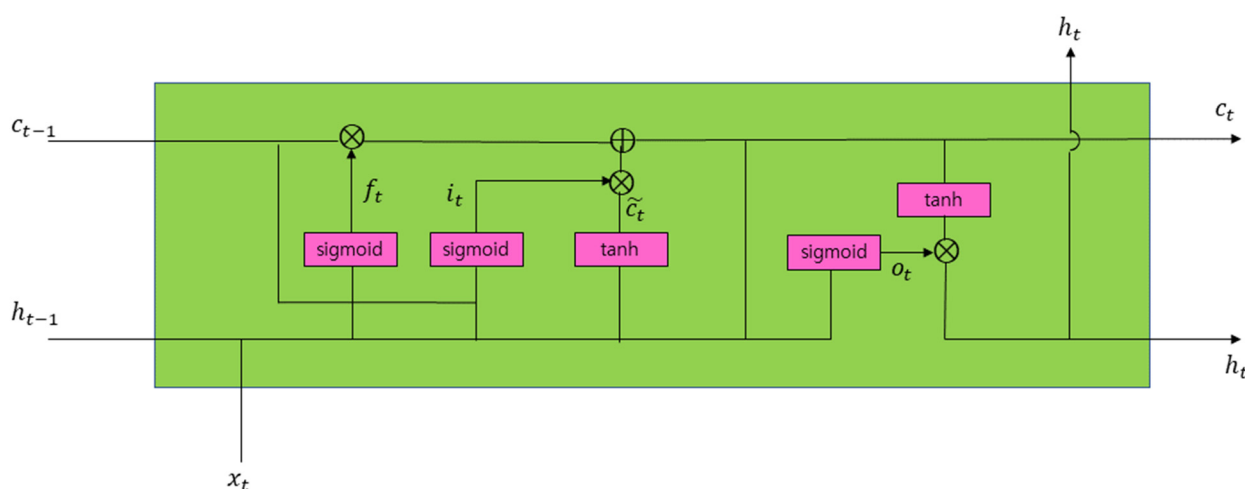


Figure 3. LSTM model [19].

2.2.3. k-Nearest Neighbor (k-NN) Model

The k-nearest neighbor algorithm is a non-parametric supervised learning method as a machine learning algorithm. The k-NN is a method of extracting the nearest k using the distance measuring metric d among the existing data when new data are given and predicting the class of new data for a classification problem and the prediction value of new data for a regression problem based on the information of the extracted data. The hyperparameters of the k-NN model include the number of searchable neighbors k and metric d . The Euclidean distance, Manhattan distance, Mahalanobis distance, correlation distance, and rank correlation distance are used as d .

2.2.4. Decision Tree Model

The decision tree model is a supervised machine learning method to solve regression or classification problems and conduct prediction by categorizing decision-making rules into a tree structure. The decision tree model is created using learning data and consists of a hierarchy of branches in which the explanatory variables are expressed as nodes and feature spaces are categorized into non-overlapping groups based on certain conditions. Each internal node represents a test on an attribute (e.g., whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules. In the final nodes at the bottom, the dependent variables are shown as the categorized groups.

2.2.5. Random Forest Model

The random forest (RF) model is an ensemble learning method as a machine learning algorithm using multiple DTs. The RF algorithm is a method for solving regression and classification problems. For classification problems, the class predicted by most DTs is produced as the output. For regression problems, the average of the predicted values of each DT is produced as the output. The RF algorithm is designed to solve the problem of overfitting in the training dataset for the DT. It divides the training dataset into different parts and randomly selects the pre-defined number of explanatory variables for each of the training dataset parts to generate a new DT.

This method reduces variance in the model by learning several DTs and averaging the prediction values from the DTs. Figure 4 is a diagram illustrating the RF algorithm.

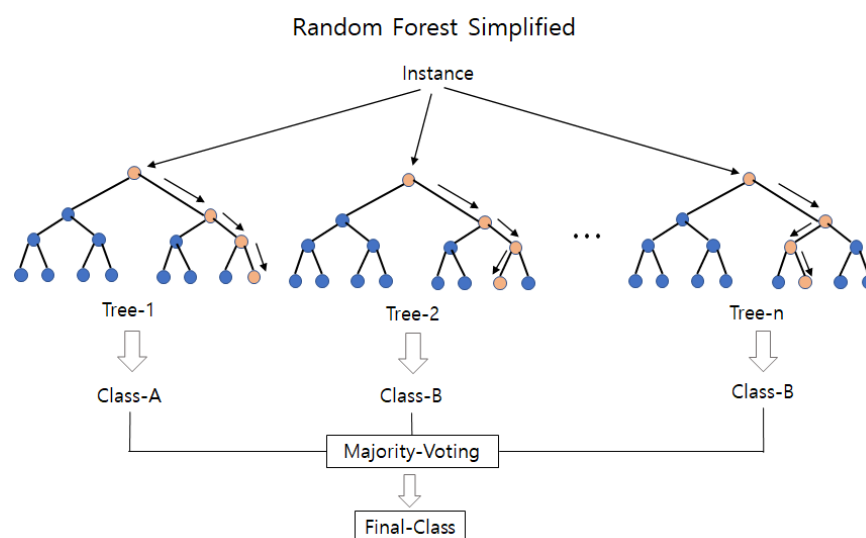


Figure 4. Diagram of random forest (source: [20]).

2.2.6. Performance Measure

As a small sample is not considered to be imbalanced data, we categorize a bankrupt company if the probability value calculated for the test dataset from 5 trained models is greater than the threshold of 0.5 and as a non-bankrupt company if it is not. As an actual non-bankrupt company can be predicted as non-bankrupt or bankrupt and the actual bankrupt company can be predicted as non-bankrupt or bankrupt, this results in the confusion matrix in Table 3.

Table 3. Confusion matrix.

		Prediction outcome	
		Non-bankrupt company	Bankrupt company
Actual outcome	Non-bankrupt company	True positive (TP)	False negative (FN)
	Bankrupt company	False positive (FP)	True negative (TN)

Accuracy, non-bankruptcy precision, non-bankruptcy recall, bankruptcy precision, bankruptcy recall, non-bankruptcy F1 score, and bankruptcy F1 score can be calculated based on the confusion matrix. By comparing the experimental results and the actual label values, we can create the graph of the ROC curve and the area under the ROC curve (AUC). The definition of each measure is provided in Equation (3).

$$\begin{aligned}
 \text{Accuracy} &= \frac{TP + TN}{TP + FP + TN + FN} \\
 \text{Non-bankruptcy Precision} &= \frac{TP}{TP + FP} \\
 \text{Non-bankruptcy Recall} &= \frac{TP}{TP + FN} \\
 \text{Bankruptcy Precision} &= \frac{TN}{FN + TN} \\
 \text{Bankruptcy Recall} &= \frac{TN}{FP + TN}
 \end{aligned} \tag{3}$$

$$\text{Non-bankruptcy F1 Score} = \frac{2 \times \text{Non-bankruptcy Precision} \times \text{Non-bankruptcy Recall}}{\text{Non-bankruptcy Precision} + \text{Non-bankruptcy Recall}}$$

$$\text{Bankruptcy F1 Score} = \frac{2 \times \text{Bankruptcy Precision} \times \text{Bankruptcy Recall}}{\text{Bankruptcy Precision} + \text{Bankruptcy Recall}}$$

The ROC AUC graph has the false positive rate as the x -axis and the true positive rate as the y -axis. It is drawn above the baseline $y = x$ and convex upward. A graph drawn above the baseline means better classification performance in the model. Hence, a higher AUC value means the model has higher classification performance.

3. Performance Analysis Results

Table 4 shows the performance results in terms of the confusion matrix, accuracy, non-bankruptcy precision, non-bankruptcy recall, bankruptcy precision, bankruptcy recall, non-bankruptcy F1 score, bankruptcy F1 score, and AUC for the LR, k-NN, DT, and RF models. The analysis was conducted for the small sample dataset created based on the 5 random samplings of 20 non-bankrupt companies, for the small sample dataset with 20 non-bankrupt companies chosen based on approximate entropy, and the total data.

The analysis of the LSTM model for the small sample data returned a confusion matrix of $\begin{bmatrix} 20 & 0 \\ 10 & 0 \end{bmatrix}$ and the model did not learn the characteristics of bankrupt companies very well as the sample of 10 companies is small. The analysis of the LSTM model for the total data returned a confusion matrix of $\begin{bmatrix} 500 & 0 \\ 10 & 0 \end{bmatrix}$ and the model did not learn the characteristics of bankrupt companies very well given the small number of 10 bankrupt companies, much like the small sample data.

Panel (a) in Table 4 shows the performance results of the models with the small dataset sampled randomly five times. The k-NN had the largest AUC, at 0.8676, and the RF had the second largest AUC, at 0.8095, indicating that the k-NN and RF outperformed the LR and DT in diagnosing non-bankrupt companies.

The LR had the highest non-bankruptcy F1 score in the performance measures of non-bankruptcy prediction and non-bankruptcy recall, while the DT had the highest bankruptcy F1 score in the performance measures of bankruptcy precision and bankruptcy recall. The LR is the best among the five models in non-bankruptcy prediction and non-bankruptcy recall, while the DT is the best in bankruptcy prediction and bankruptcy recall.

Panel (b) in Table 4 provides the performance measure results of the models for the small sample data sampled using approximate entropy. Here, RF had the highest AUC, at 0.8326, and the k-NN had the second highest AUC, at 0.8076. Further, RF and the k-NN outperformed the LR and DT in diagnosing non-bankrupt companies. RF had the best performance in terms of non-bankruptcy precision, non-bankruptcy recall, bankruptcy precision, and bankruptcy recall. Thus, RF is the best of the five models in terms of precision and recall.

The performance results in Panels (a) and (b) in Table 4 show that the AUC of the k-NN and RF was higher than that of the LR and DT. The analysis results for the small dataset sampled using approximate entropy showed better performance measures of bankruptcy precision and bankruptcy recall than those for the small sample data sampled randomly.

Panel (c) in Table 4 shows the performance results of the models for the total data. Like the LSTM, the small number of 10 bankrupt companies led to overfitting to non-bankrupt companies and subsequently resulted in 0 or close to 0 in bankruptcy precision and bankruptcy recall, despite generating samples for 90 bankrupt companies, which is 9 times larger than the input data. These models did not predict bankruptcies very well.

Table 4. Performance results: LR, k-NN, DT, and RF models for the small sample and total dataset.

Model	Confusion Matrix	Accuracy	Non-Bankruptcy Precision	Non-Bankruptcy Recall	Bankruptcy Precision	Bankruptcy Recall	Non-Bankruptcy F1 Score	Bankruptcy F1 Score	AUC
LR	$\begin{bmatrix} 183 & 23 \\ 49 & 15 \end{bmatrix}$	0.7333	0.7888	0.8883	0.3947	0.2344	0.8356	0.2941	0.7457
k-NN	$\begin{bmatrix} 183 & 23 \\ 53 & 11 \end{bmatrix}$	0.7185	0.7754	0.8883	0.3235	0.1719	0.8281	0.2245	0.8676
DT	$\begin{bmatrix} 155 & 51 \\ 42 & 22 \end{bmatrix}$	0.6556	0.7868	0.7524	0.3014	0.3438	0.7692	0.3212	0.6587
RF	$\begin{bmatrix} 178 & 28 \\ 49 & 15 \end{bmatrix}$	0.7148	0.7841	0.8641	0.3488	0.2344	0.8222	0.2804	0.8095
(a) Performance results of the LR, k-NN, DT, and RF for the small random sample									
Model	Confusion Matrix	Accuracy	Non-Bankruptcy Precision	Non-Bankruptcy Recall	Bankruptcy Precision	Bankruptcy Recall	Non-Bankruptcy F1 Score	Bankruptcy F1 Score	AUC
LR	$\begin{bmatrix} 187 & 9 \\ 58 & 16 \end{bmatrix}$	0.7519	0.7633	0.9541	0.64	0.2162	0.8481	0.3232	0.7929
k-NN	$\begin{bmatrix} 187 & 9 \\ 53 & 21 \end{bmatrix}$	0.7704	0.7792	0.9541	0.7	0.2838	0.8578	0.4038	0.8076
DT	$\begin{bmatrix} 168 & 28 \\ 41 & 33 \end{bmatrix}$	0.7444	0.8038	0.8571	0.5410	0.4459	0.8296	0.4889	0.5979
RF	$\begin{bmatrix} 182 & 14 \\ 43 & 31 \end{bmatrix}$	0.7889	0.8089	0.9286	0.6889	0.4189	0.8646	0.5210	0.8326
(b) Performance results: LR, k-NN, DT, and RF for the small dataset sampled using approximate entropy									
Model	Confusion Matrix	Accuracy	Non-Bankruptcy Precision	Non-Bankruptcy Recall	Bankruptcy Precision	Bankruptcy Recall	Non-Bankruptcy F1 Score	Bankruptcy F1 Score	AUC_1 AUC_2
LR	$\begin{bmatrix} 4514 & 2 \\ 74 & 0 \end{bmatrix}$	0.9834	0.9839	0.9995	0	0	0.9916	Not defined	0.7906 0.9998
k-NN	$\begin{bmatrix} 4503 & 13 \\ 74 & 0 \end{bmatrix}$	0.9810	0.9839	0.9971	0	0	0.9904	Not defined	0.6500 0.9998
DT	$\begin{bmatrix} 4398 & 118 \\ 68 & 6 \end{bmatrix}$	0.9595	0.9848	0.9739	0.0484	0.0811	0.9793	0.0606	0.5477 0.9998
RF	$\begin{bmatrix} 4514 & 2 \\ 74 & 0 \end{bmatrix}$	0.9834	0.9839	0.9996	0	0	0.9917	Not defined	0.7389 0.9998

Table 4. *Cont.*

(c) Performance results: LR, k-NN, DT, and RF for the total data, where AUC_1 = AUC at the threshold of 0.5 and AUC_2 = AUC using the optimal threshold.					
	Accuracy	Precision	Recall	F1 Score	AUC
Logistic	0.7570	0.0001	0.2174	0.0002	0.4872
SVM	0.7236	0.0002	0.3913	0.0003	0.5575
RF	0.9899	0.0023	0.2174	0.0045	0.6037
RNN	0.9789	0.0024	0.4783	0.0048	0.7286
LSTM	0.9936	0.0058	0.3478	0.0114	0.6707
Ensemble	0.9826	0.0029	0.4783	0.0058	0.7305
(d) Performance results: Bankruptcy Forecasting Performance by Methodology [7]					

To overcome the problem of worsening performance due to data imbalance, we propose a method to improve the AUC performance measure by applying the optimal threshold. The LR, k-NN, DT, and RF models predict bankruptcy in the $(n + 1)$ -th year based on n -th year's financial information. The anomaly score is the difference between the value indicating bankruptcy by the model in year $(n + 1)$ and the value indicating actual bankruptcy (bankruptcy is 1, non-bankruptcy is 0). A higher anomaly score means the status in year $(n + 1)$ is likely to be an anomaly (bankruptcy). We use a supervised thresholding mechanism to detect the optimal threshold for binarizing points as anomalous or not. We obtain the minimum and maximum values of the anomaly scores calculated on the training dataset and then consider each value from the minimum score to the maximum score with a small step and select the value returning the highest F-score as the overall model threshold. Figure 5 shows the optimal threshold mechanism.

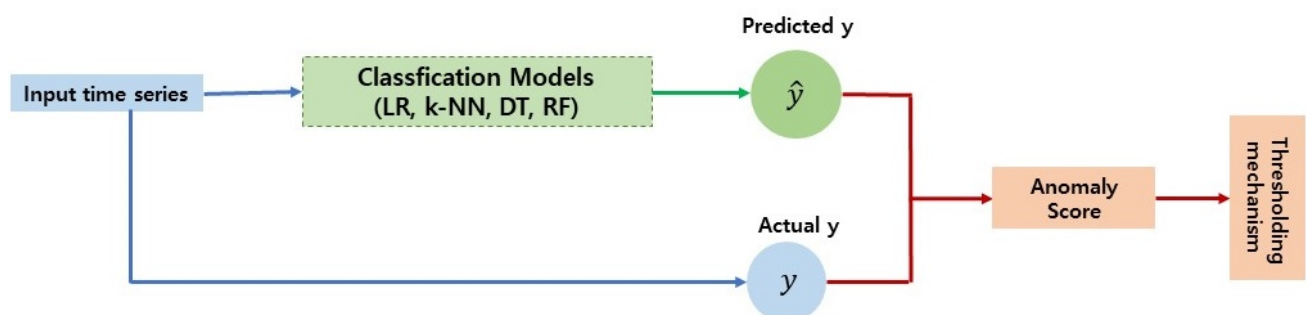


Figure 5. Optimal threshold mechanism.

We believe that our proposed method would greatly enhance the value of the AUC as in Panel (c) in Table 4, where the optimal threshold is applied. Using the information obtained from the anomaly scores from the training dataset, we use the AUC-based thresholds mechanism to identify the optimal thresholds for determining bankruptcy or non-bankruptcy. The anomaly score is measured for the test dataset using the identified optimal thresholds. If the anomaly score is above the identified threshold, then we regard it as a bankruptcy (1). If it is below the identified threshold, then we regard it as non-bankrupt (0).

As shown in Table 4, the AUC of the performance measures in Panels (a) and (b) was generally larger than that in Panel (c), and the performance measures of bankruptcy precision and bankruptcy recall were significant, suggesting that the method of learning and predicting bankruptcy by configuring only some non-bankrupt companies as the training dataset when there are few bankrupt companies can create a model that classifies bankrupt and non-bankrupt companies well without greatly sacrificing accuracy. However, the AUC value when the optimal threshold is applied in Panel (c) in Table 4 is greater than the AUC of the performance measures in Panels (a) and (b) in Table 4, which suggests that identifying the optimal threshold based on the AUC for classification is more effective than applying small sampling to imbalanced data.

The observed difference in performance according to the models suggests that the RF is the best model, with the highest AUC of 0.8326 in Panel (b) in Table 4 to diagnose non-bankrupt companies. It is also the best model, with the highest non-bankruptcy F1 and bankruptcy F1 scores of 0.8646 and 0.5210, respectively, in Panel (b), to classify non-bankrupt and bankrupt companies. Panel (d) in Table 4 is the result of [7], an analysis of monthly financial information of about 45,472 non-financial firms collected from January 2007 to December 2019 from the Compustat North America Dataset. Since it included the monthly financial information of 2057 bankrupt companies, it offers a sufficient training dataset for the RNN and LSTM. Hence, the RNN and LSTM had significant AUC performance.

The performance measures of precision, recall, F1 score, and AUC from the LR and RF models trained with the small random sample and the LR and RF models trained with the small sample created using approximate entropy were higher than those of the LR and RF models in Panel (d) in Table 4. In addition, the AUC performance measure in Panel (c) and the AUC performance measure when the optimal threshold was applied were higher than those of the LR and RF models in Panel (d).

These results suggest that we identified a data processing method and two data sampling methods that improve the performance of the corporate bankruptcy prediction models for imbalanced corporate financial data, and a method using the optimal threshold to improve AUC performance even when the total imbalanced data are used.

Figure 6 depicts four ROC AUC curves for four small random samples where a small random sample is obtained by random sampling of the non-bankrupt companies five times. From Figure 6, it can be seen that the performance of diagnosing non-bankrupt companies is excellent for k-NN and RF, and it is difficult to compare the performances of LR and DT. Figure 7 plots the ROC AUC curve the small sample generated using approximate entropy, where RF had the best performance in diagnosing non-bankrupt companies, followed by k-NN, LR, and DT. These results are consistent with those in Panels (a) and (b) in Table 4. Figure 8 represents the optimal threshold identified based on the AUC for each model as a dot.

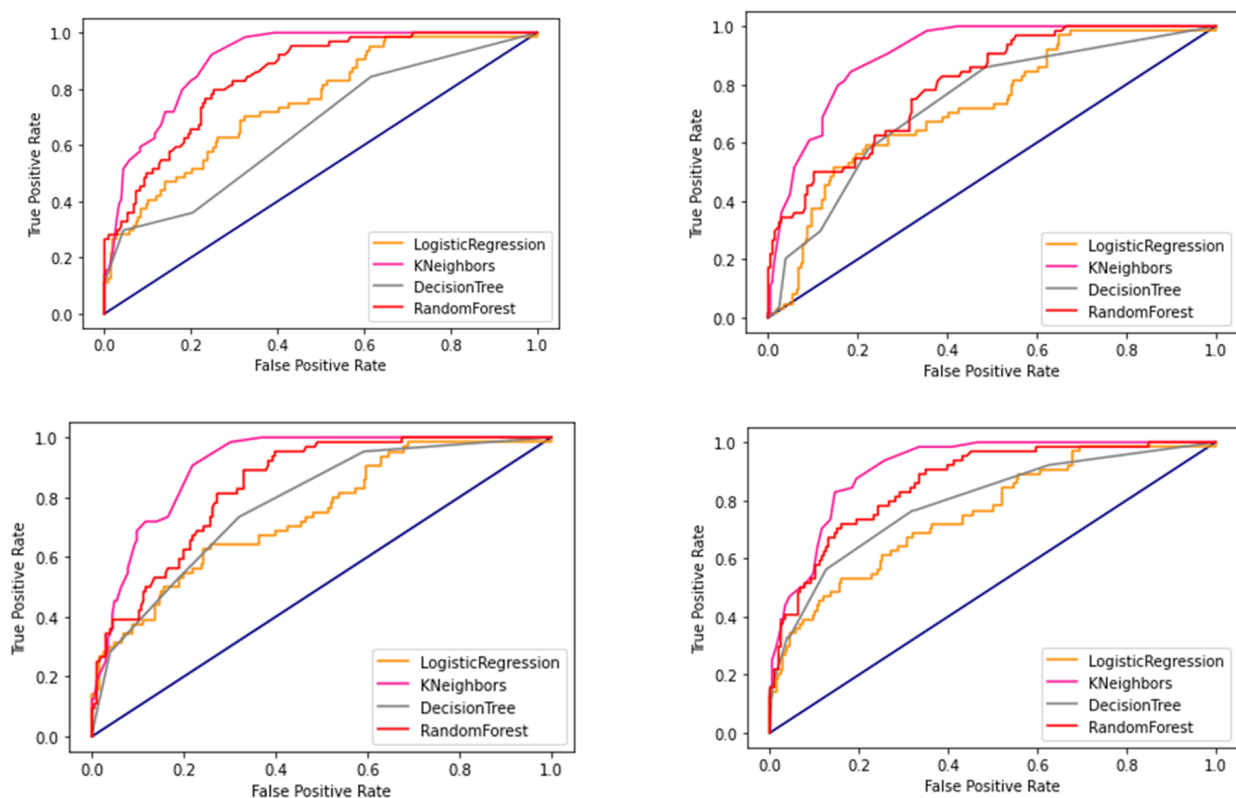


Figure 6. ROC AUC graph for four small data samples sampled randomly.

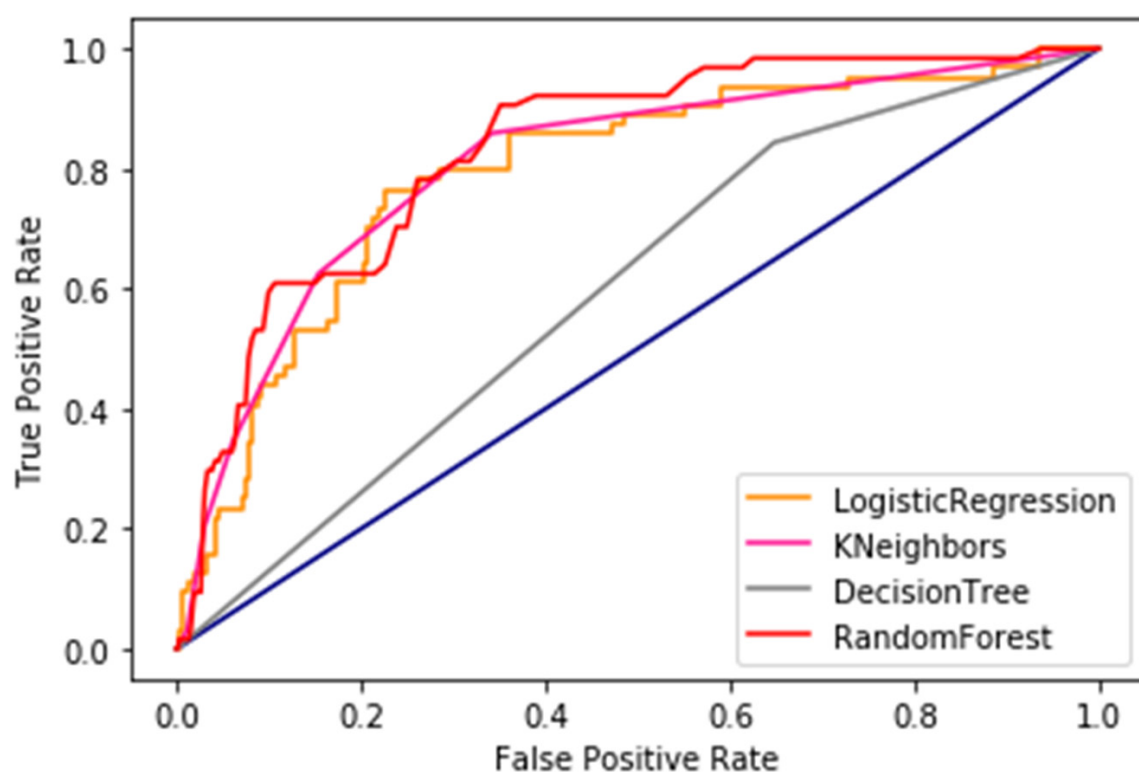


Figure 7. ROC AUC graph for the small sample created using approximate entropy.

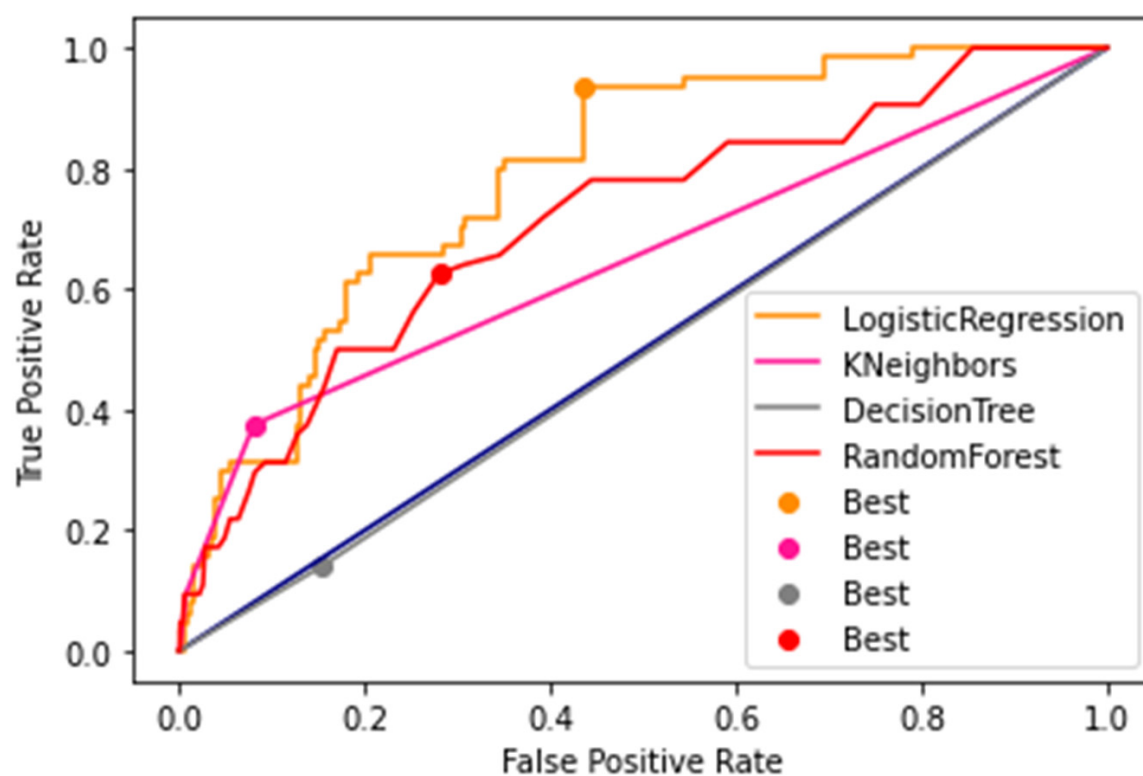


Figure 8. ROC AUC curve with the highest score (optimal threshold = 0.4357 for the LR, 0.0800 for k-NN, 0.1326 for DT, and 0.3789 for RF).

4. Conclusions

We collected 13 financial information items for about 1020 companies listed on KOSPI and KOSDAQ from 2012 to 2021 and compared the bankruptcy prediction performance of the LSTM, LR, k-NN, DT, and RF models.

As the collected financial data on the companies in Korea represents a small sample due to the small number of bankruptcies, the LSTM is not a suitable corporate bankruptcy prediction model.

As the LR, DT, RF, and k-NN classification models do not use time series data as an input, we configured financial information in year n and bankruptcy or non-bankruptcy in year $n + 1$ as the input and label data, which made it possible to generate nine times more input data compared to the LSTM model and resulted in higher bankruptcy prediction performance.

The analysis of small sample data compared the random sampling method of non-bankrupt companies in the training dataset with the sampling method using approximate entropy, where the latter outperformed the former method in bankruptcy precision and bankruptcy recall. The performance measures of the models trained with the dataset constructed using five random samples and those generated using approximate entropy were higher than those of the models trained with the sample data identified from [7], confirming that our proposed method can improve the bankruptcy prediction performance using a small sample relative to a large sample for an imbalanced dataset. In particular, for the unbalanced financial data collected in this study, our analysis proved that the AUC performance measure can be greatly improved by classification using the optimal threshold identified based on the AUC, as shown in Panel (c) in Table 4. The results of this study provide useful information for selecting a suitable bankruptcy prediction model when using a dataset with relatively few bankrupt companies.

In this study, the number of bankrupt companies for which financial information was collected is very small, so we need to use a strategy such as the cross-validation technique to reduce the variation that may occur when performing experiments by sampling non-bankrupt companies. We plan to conduct experiments applying these techniques in future works. We measured the performance by averaging the experimental results of random sampling with five repeated experiments, and it is necessary to increase the number of such repeated experiments to improve the reliability of the experimental results, so we plan to do so in future works. In this study, we did not conduct an experiment to compare the experimental results of the benchmark model and benchmark dataset with the results of this study, so we plan to add them in future works to ensure the reliability of the research results.

Funding: This research received no external funding.

Data Availability Statement: The code and datasets used and analyzed during this study can be obtained from the online resource (<https://github.com/shnoh92/Corporate-Bankruptcy-Prediction-Models-Based-on-Imbalanced-Financial-Data/> accessed on 5 March 2023).

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table A1. Twenty-nine variables that correspond to the criteria characterizing bankrupt and non-bankrupt companies.

Category	Section	Feature
Financial Statements	Balance Sheet (1000 won)	Accumulations
		Retained Earnings
		Net assets of controlling shareholders (before capital stock reduction)
		Owners of Parent Equity
		Total Equity
	Comprehensive Income Statement (1000 won)	Earnings before tax
		(Total Comprehensive Income Attributable to) Owners of Parent Equity
		Total Comprehensive Income
	Cash Flow Statement (1000 won)	Cash Flow
Financial Ratio	Stability (%)	Intangible Asset Ratio
		Equity Capital Ratio
		Borrowings and Bonds Payable Ratio
		Borrowed Capital Ratio
		Cash Flow/Total Debt
		Cash Flow/Total Equity
		Cash Flow/Total Asset
	Growth (yearly) (%)	Total Asset Growth Rate
	Profitability (%)	Operating Revenue/Operating Expense
		Profit Margin Ratio
		ROA (Current Net Income)
		ROA (Earnings before tax)
		ROA (Operating Profit)
		ROA (Total Comprehensive Income)
		ROE (Current Net Income)
		ROE (Earnings before tax)
		ROE (Operating Profit)
		ROE (Net profit of controlling shareholders)
	Activity (times)	Total Debt Turnover
		Total Asset Turnover

References

- Oh, W.S.; Kim, J.H. Forecasting corporate bankruptcy with artificial intelligence. *J. Ind. Converg.* **2017**, *15*, 17–32.
- Cha, S.; Kang, J. Corporate default prediction model using deep learning time series algorithm, RNN and LSTM. *J. Intell. Inf. Syst.* **2018**, *24*, 1–32.
- Barboza, F.; Kimura, H.; Altman, E. Machine learning models and bankruptcy prediction. *Expert Syst. Appl.* **2017**, *83*, 405–417. [[CrossRef](#)]
- Falavigna, G. Financial ratings with scarce information: A neural network approach. *Expert Syst. Appl.* **2012**, *39*, 1784–1792. [[CrossRef](#)]
- Hinton, G.E.; Osindero, S.; The, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [[CrossRef](#)] [[PubMed](#)]
- Jang, Y.; Jeong, I.; Cho, Y.; Ahn, H. Business Failure Prediction with LSTM RNN in the Construction Industry. In Proceedings of the ASCE 2019 International Conference on Computing in Civil Engineering, Atlanta, GA, USA, 17–19 June 2019; pp. 1–8.

7. Kim, H.; Cho, H.; Ryu, D. Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. *Comput. Econ.* **2022**, *59*, 1231–1249. [[CrossRef](#)]
8. Odom, M.D.; Sharda, R. A neural network model for bankruptcy prediction. In Proceedings of the 1990 IJCNN International Joint Conference on Neural Networks, San Diego, CA, USA, 17–21 June 1990; pp. 163–168.
9. Wilson, R.L.; Sharda, R. Bankruptcy prediction using neural networks. *Decis. Support. Syst.* **1994**, *11*, 545–557. [[CrossRef](#)]
10. Kim, H.; Cho, H.; Ryu, D. Corporate default predictions using machine learning: Literature review. *Sustainability* **2021**, *12*, 6325. [[CrossRef](#)]
11. Brygata, M. Consumer Bankruptcy Prediction Using Balanced and Imbalanced Data. *Risks* **2022**, *10*, 24. [[CrossRef](#)]
12. Zhou, L. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowl. Based Syst.* **2013**, *41*, 16–25. [[CrossRef](#)]
13. Garcia, V.; Jose, S.S.; Ramon, A.M. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowl. Based Syst.* **2012**, *25*, 13–21. [[CrossRef](#)]
14. Syed, N.; Sharifah, H.; Shafinar, I.; Bee, W.Y. Personal bankruptcy prediction using decision tree model. *J. Econ. Financ. Adm. Sci.* **2019**, *24*, 157–170. [[CrossRef](#)]
15. Amidon, A. PyOD: A Unified Python Library for Anomaly Detection. 11 May 2021. Available online: <https://towardsdatascience.com/pyod-a-unified-python-library-for-anomaly-detection-3608ec1fe321> (accessed on 15 January 2023).
16. Mishra, S.; Kshisagar, V.; Dwivedula, R.; Hota, C. Attention-Based Bi-LSTM for Anomaly Detection on Time-Series Data. In Proceedings of the 2021 ICANN International Conference on Artificial Neural Networks, Online, 14–17 September 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 129–140.
17. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the 2021 NeurIPS 35th Conference on Neural Information Processing Systems, online, 6–14 December 2021; pp. 1–14.
18. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **2000**, *278*, H2039–H2049. [[CrossRef](#)] [[PubMed](#)]
19. Noh, S.-H. Analysis of gradient vanishing of RNNs and performance comparison. *Information* **2021**, *12*, 442. [[CrossRef](#)]
20. Jagannath, V. Random Forest Template Tibco Spotfirer Wiki Page. 24 March 2017. Available online: <https://community.tibco.com/wiki/random-forest-template-tibco-spotfirer-wiki-page> (accessed on 15 January 2023).

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