sustainability

MDPI

Review

Corporate Default Predictions Using Machine Learning: Literature Review

Hyeongjun Kim [1], Hoon Cho [2] and Doojin Ryu [3,*]

[1] Department of Business Administration, Yeungnam University, Gyeongsan 38541, Korea; hkim@yu.ac.kr
[2] College of Business, Korea Advanced Institute of Science and Technology, Seoul 02455, Korea; hooncho@kaist.ac.kr
[3] College of Economics, Sungkyunkwan University, Seoul 03063, Korea
* Correspondence: doojin.ryu@gmail.com

Received: 22 July 2020; Accepted: 31 July 2020; Published: 6 August 2020

Abstract: Corporate default predictions play an essential role in each sector of the economy, as highlighted by the global financial crisis and the increase in credit risk. This study reviews the corporate default prediction literature from the perspectives of financial engineering and machine learning. We define three generations of statistical models: discriminant analyses, binary response models, and hazard models. In addition, we introduce three representative machine learning methodologies: support vector machines, decision trees, and artificial neural network algorithms. For both the statistical models and machine learning methodologies, we identify the key studies used in corporate default prediction. By comparing these methods with findings from the interdisciplinary literature, our review suggests some new tasks in the field of machine learning for predicting corporate defaults. First, a corporate default prediction model should be a multi-period model in which future outcomes are affected by past decisions. Second, the stock price and the corporate value determined by the stock market are important factors to use in default predictions. Finally, a corporate default prediction model should be able to suggest the cause of default.

Keywords: classification; default prediction; financial engineering; forecasting; machine learning

JEL Classification: G10; G17; G33

1. Introduction

Forecasts of corporate defaults are used in various fields across the economy. Corporations can diagnose their current statuses based on prediction models and establish their strategies. Executives can run their businesses more stably by managing key indicators that affect corporate default risk. Investors can revise their strategies and improve their portfolios by examining the likelihood of corporate defaults. Additionally, governments can establish macroprudential policies and improve related financial regulations using corporate default predictions. In these ways, default prediction models help in designing and improving the financial system. Moreover, by employing machine learning algorithms and statistical models, corporate default predictions are at the cutting edge of advanced financial engineering. The recent global financial crisis and the increase in credit risk highlight the importance of this field. Because of their importance, corporate default predictions have been extensively studied since the work of Beaver [1].

Thus far, several structural models have been used to explain corporate defaults. Merton [2] develops the "distance-to-default" measure of corporate default risk using a Black-Scholes-type pricing model. In a recent study, Jessen and Lando [3] confirm that the distance-to-default measure can robustly detect corporate default risk. They also present a distance-to-default measure with an adjustment

using the stochastic volatility of the value of assets. Glover [4] proposes a structural model to calculate a corporation's expected default costs. Brogaard et al. [5] use the distance-to-default approach to show that enhanced stock liquidity reduces corporate default risk. Hillegeist et al. [6] argue that a market-based measure based on the Black–Scholes–Merton model performs better than Altman's [7] Z-score and Ohlson's [8] O-score when assessing a discrete hazard model. Duffie et al. [9] propose a doubly stochastic model to estimate the term structure of corporate default risk. Recently, however, reduced-form models with statistical approaches and machine learning algorithms that predict the likelihood of a corporate default provide more satisfactory results than structural models in general.

In this study, we review the corporate default prediction literature from the perspectives of financial engineering and machine learning simultaneously. We attempt to identify new opportunities in the field of machine learning for predicting corporate defaults by comparing these kinds of literature. This study therefore examines the research on corporate default forecasts thus far and introduces representative methodologies and major studies by categorizing statistical approaches and machine learning techniques. We define three generations of the main statistical models for corporate default forecasting as discriminant analyses, binary response models, and hazard models, and then we study each generation. Among machine learning algorithms, classification methodologies are mainly used in this setting, and support vector machines (SVMs), decision trees, and artificial neural network algorithms are typical. There are already several studies reviewing the field of bankruptcy prediction [10–13]. However, we will focus on discovering new research topics and challenges in this study. The development of machine learning methodologies is expected to further accelerate innovation in the financial sector, leading to the emergence of new financial services and the rise of the data economy based on the distribution of data.

Before starting the review, we need to set the scope of the investigation. In this study, we define the financial literature as academic journals covering the following areas: accounting, finance, financial economics, economics, and econometrics. The machine learning literature includes academic papers covering computer science, statistics, and operations research. We search the databases, Google Scholar and EBSCOhost, to exhaustively collect the related literature. The keywords used for the search are "corporate default prediction," or "bankruptcy prediction." We also include the keyword "machine learning" when we find the machine learning literature. After we construct our own classification system, we use the name of each methodology as additional keywords. The citation count (e.g., Journal Citation Reports–Clarivate Analytics) is mainly used for thesis selection, but our subjective evaluations are also considered.

This study explores various techniques and algorithms used for corporate default predictions. The remainder of the paper is organized as follows. Section 2 focuses on statistical approaches for corporate default predictions, and Section 3 reviews several machine learning techniques. Section 4 concludes the study.

2. Corporate Default Prediction Using Statistical Approaches

Corporate default forecasts using statistical models can be largely classified into three generations. Table 1 shows these three primary methodologies and their representative studies. Various studies have been conducted as each methodology has expanded, and these methods are actively used to this day.

Table 1. Classification of statistical models for corporate default prediction.

Statistical Approaches	References
Discriminant Analyses	Beaver (1966); Altman (1968); Mare et al. (2017)
Binary Response Models	Ohlson (1980); Zmijewski (1984); Foreman (2003); Campbell et al. (2008); Kukuk and Rönnberg (2013); Aretz et al. (2018)
Hazard Models	Shumway (2001); Chava and Jarrow (2004); Nam et al. (2008); Bonfim (2009); Dakovic et al. (2010); Duan et al. (2012); Figlewski et al. (2012); Tian et al. (2015); Traczynski (2017)

2.1. Discriminant Analysis

Studies in the first generation of the corporate default literature use discriminant analysis. Discriminant analysis is a default prediction methodology that has been widely used since the work of Beaver [1] and Altman [7]. These studies build reduced-form default prediction models using discriminant analysis and provide ordinal rankings of default risk by generating credit scores. The famous Altman Z-score is one example, and subsequent studies use this methodology [14,15]. Discriminant analysis selects the variables that can best determine whether a company is bankrupt and calculates the discriminant function as a linear combination of these variables, as follows:

$$D = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m, \tag{1}$$

where D is the discriminant score calculated by the discriminant function, β_0 is a constant, β_m is an estimated coefficient, and X_m is an explanatory variable. An observation is classified as normal if the discrimination score is below a certain threshold and is classified as being in default if the score is above that threshold. However, discriminant analysis requires the assumption that the independent variables follow multivariate normal distributions, the covariance matrix between the two groups is defined as normal, and defaults are identical [16]. The discriminant analysis does have the advantage that corporations can be ranked according to their degree of default risk.

2.2. Binary Response Models

The second generation of corporate default predictions uses binary response models. A binary response model is a model that defines a corporation's state as either normal (= 0) or in default (= 1). It estimates the probability of a default using explanatory variables and applies a logistic or probit function in most cases. A representative example is the O-score with a logistic function, introduced by Ohlson [8], whereas Zmijewski [17] tests corporate default risk using a probit model. These binary models allow bankruptcy probabilities to be calculated over the next period. As Foreman [18] and Charitou et al. [16] explain, the binary response model defines the probability that corporation $n = 1, \ldots, N$ defaults, P_n, as follows:

$$P_n(y_n = 1) = 1/(1 + e^{-z}) = 1/\{1 + \exp[\,-(\beta_0 + \beta_{1\times 1,n} + \beta_2 X_{2,n} + \ldots + \beta_m X_{m,n})]\}, \tag{2}$$

where $y_n = 1$ if corporation n defaults and 0 otherwise. $P_n(y_n = 1)$ is the probability of default for corporation n, $\beta_1, \beta_2, \ldots, \beta_m$ are slope coefficients, and $X_{1,n}, X_{2,n}, \ldots, X_{m,n}$ are explanatory variables for corporation n.

The likelihood function is given by

$$L = \prod_{n=1}^{N} F(\beta' X_n)^{y_n} (1 - F(\beta' X_n))^{1-y_n}, \tag{3}$$

where $F(\beta' X_n) = 1/\{1 + \exp[-(\beta_0 + \beta_1 X_{1,n} + \beta_2 X_{2,n} + \ldots + \beta_m X_{m,n})]\}$. The coefficients can be estimated using the maximum likelihood technique.

A binary response model has several advantages over discriminant analysis for corporate default forecasting. First, a binary response model does not require any assumptions about the probability of default or the distributions of the predictor variables. Second, it can test the significance of individual independent variables. Lastly, it can be used to calculate the probability of default in the next period. Campbell et al. [19] extend the binary response model by investigating corporate defaults using a multiple logit model. This approach allows us to calculate corporate default probabilities and predict default risk over several periods. Aretz et al. [20] adopt Campbell et al.'s [19] methodology and use data for non-U.S. companies to identify a significantly positive default risk premium. Kukuk and Rönnberg [21] extend the binary response model by proposing a mixed logit model that allows stochastic parameters and non-linearities in the regressor variables. Bonfim [22] tests macroeconomic and financial data using a probit model and argues that corporate defaults are driven by multiple firm-specific factors. However, the results show that macroeconomic factors are also important in estimating default risk over time.

2.3. Hazard Models

The third generation of corporate default predictions includes studies using hazard models. Shumway [23] uses a duration analysis with a hazard model and shows that this approach predicts corporate defaults better than traditional single-period models do. The hazard model is also referred to as survival analysis and can be used to calculate the probability of a corporate default over time. This default prediction methodology uses Cox's [24] hazard regression model by defining a corporation's status as either normal (= 0) or in default (= 1). The corporation's status is no longer observed once a default event occurs. Assuming that T is the time at which a company defaults, company's survival function at time t can be expressed as follows:

$$S(t) = Pr(T \geq t) \tag{4}$$

The hazard function, $\lambda(t)$, indicates the instantaneous failure rate at time t and is defined as follows:

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{5}$$

The Cox proportional hazard model defines the instantaneous failure rate, $\lambda(t)$, as being proportional to the unspecified baseline hazard rate, $\lambda_0\,(t)$, as follows:

$$\lambda(t|X_n) = \lambda_0(t) exp(\beta' X_n) \tag{6}$$

where β is a column vector composed of regression coefficients, and X_n is the set of explanatory variables for the nth company. The Cox model is a semi-parametric model consisting of a non-parametric base risk rate $\lambda_0(t)$ and a parametric factor $exp(\beta' X_n)$. The partial likelihood function for the regression coefficient β is calculated as follows:

$$PL(\beta) = \prod_{i=1}^{N} \left[\frac{exp(\beta' X_i)}{\sum_{j=1}^{N} Y_{ij} exp(\beta' X_i)} \right]^{\delta_i} \tag{7}$$

where Y_{ij} is an indicator variable equal to one if $t_j \geq t_i$ and zero otherwise. δ_i is an indicator variable equal to one when an observation is not censored and zero otherwise. The parameters are estimated such that the partial likelihood function is maximized.

The hazard model is developed further by many subsequent studies. Chava and Jarrow [25] confirm that the hazard model exhibits superior prediction performance, and they address the importance of industry effects and market variables. Nam et al. [26] extend Shumway's [23] analysis by including time-varying covariates to incorporate macroeconomic dependencies. Dakovic et al. [27]

fit a generalized linear mixed model, including unobserved heterogeneity between industry sectors, into a discrete hazard model, and show that the new model outperforms conventional models with Altman's [7] variables. Duan, Sun, and Wang [28] develop a forward intensity model from the hazard model and predict corporate default probabilities over multiple periods. Traczynski [29] extends the hazard model by taking a Bayesian model-averaging approach and shows that this approach leads to a better prediction performance relative to other typical models. Figlewski et al. [30] evaluate the effect of macroeconomic conditions on corporate default risk using reduced-form Cox intensity models. Tian et al. [31] use a variable selection technique with a discrete hazard model and demonstrate that the variables selected using the least absolute shrinkage and selection operator can improve prediction performance.

3. Corporate Default Prediction Using Machine Learning Techniques

Samuel [32] proposes the concept of machine learning and defines it as "a discipline that gives computers the ability to learn without a clear program". Mitchell [33] further develops this notion of machine learning, saying, "a computer program is said to *learn* from experience E with respect to some class of tasks T and performance measures P, if its performance at tasks in T, as measured by P, improves with experience." In this regard, a machine learning algorithm for corporate default prediction can be described as a series of processes that improve the default indicator (P) to perform the task of predicting corporate credit risk (T) using actual corporate credit information (E). Researches on corporate default prediction using machine learning techniques are conducted in various ways, especially in the field of computer science, and Barboza et al. [34] argue that machine learning models exhibit better performances in predicting the corporate bankruptcy. Table 2 shows three important methodologies and their representative studies. In most cases, a default prediction using machine learning adopts a classification problem that categorizes a company's status as being in one of two or more states, defined as normal (= 0) and in default (= 1), and it calculates the probability that the company is in a specific state. Thus, machine learning algorithms for solving classification problems are primarily used; representative examples include SVMs, decision trees, and artificial neural network algorithms.

Table 2. Classification of machine learning techniques for corporate default prediction.

Machine Learning Techniques	References
Support Vector Machines	Shin et al. (2005); Chen (2011); Lu et al. (2014)
Decision Trees	Olson et al. (2012); Tsai et al. (2014)
Artificial Neural Networks	Yang, Platt, and Platt (1999); Falavigna (2012); Iturriaga and Sanz (2015); Azayite and Achchab (2016)

3.1. Support Vector Machines

The SVM classification algorithm is widely used in various fields, including corporate default predictions. This algorithm uses a separating hyperplane to classify n observations, each of which has p features. Each observation can be classified as having one of two statuses, defined as $y_i \in \{-1,1\}$, where y_i represents the status of the ith observation. The SVM algorithm needs to determine the farthest separating hyperplane, and it allows some misclassifications to avoid the overfitting problem. The optimal separating hyperplane can be represented by the following equation

$$\underset{\beta_0,\beta_1,\cdots,\beta_p,\varepsilon_0,\varepsilon_1,\cdots,\varepsilon_n}{max} M \tag{8}$$

subject to $\sum_{j=1}^{p} \beta_j^2 = 1$, $y_i\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) \geq M(1-\varepsilon_i)$, $\varepsilon_i \geq 0$, $\sum_{i=1}^{n} \varepsilon_i \leq C$, where β_0, β_1, $\cdots$, β_p are hyperplane parameters and ε_0, ε_1, $\cdots$, ε_n are slack variables to allow for some misclassified observations. M denotes the margin, which means the distance between the separating hyperplane

and the observations. The tuning parameter, C, is greater than zero and determines the limits of the classification errors [35].

Shin et al. [36] apply an SVM to corporate default prediction and show that the SVM performs better than a back-propagation neural network model. Chen [37] compares several default prediction models and claims that the SVM has high accuracy and performs well for short- and long-term default predictions. Liang et al. [38] also show that the SVM yields the best performance when predicting bankruptcy using financial ratios and corporate governance indicators. Lu et al. [39] extend the SVM methodology using hybrid switching particle swarm optimization.

3.2. Decision Trees and Random Forests

The decision tree (DT) algorithm is a methodology for solving regression or classification problems by charting decision rules in a tree structure. The DT algorithm divides a feature space, composed of a combination of p explanatory variables $X_1, X_2, \ldots, X_p$, into J non-overlapping regions $R_1, R_2, \ldots, R_J$, and it makes the same prediction for observations belonging to the same domain. The following *Gini index* measures the quality of separation

$$G = \sum_{j=1}^{J} \hat{p}_{mj}\left(1 - \hat{p}_{mj}\right) \tag{9}$$

where J is the number of states and $\hat{p}_{mj}$ is the proportion of state j in region R_m [35]. The *pruning* process, which establishes individual decision trees, is performed using the Gini index. The DT algorithm has the advantage that the model is intuitive and easy to interpret, but it has the limitation that the over-fitting problem is likely to occur in the process of dividing the feature space or producing branches, and the prediction accuracy is reduced as a result.

The random forest (RF) algorithm is a machine learning algorithm that uses multiple DTs. The RF algorithm chooses a pre-determined number of explanatory variables through randomization when it creates a new DT. Generally, the number of variables selected is given by the square root of p, the total number of explanatory variables. If we denote this number as k, then the RF algorithm generates multiple DTs with k randomly selected explanatory variables. For a classification problem, this model's prediction is based on the most commonly predicted result across the DTs. Olson et al. [40] argue that the DT algorithm is more understandable and accurate than other machine learning algorithms. Tsai et al. [41] compare DT ensembles with SVMs and multilayer perception neural networks and claim that DT ensembles perform the best. Zięba et al. [42] extend the DT algorithms by using the extreme gradient boosting and synthetic features generation. Jardin [43] notably proposes a corporate default prediction model with ensembles of Kohonen maps and argues that this approach is highly efficient.

3.3. Artificial Neural Networks

An artificial neural network is a machine learning algorithm devised according to the process by which real human brains operate. This algorithm solves complex problems by mimicking the structure of the brain and connecting artificial neurons using simple structures. Neurons, which are the basic units making up the human brain and spinal cord, are responsible for transmitting the signals that they receive to other connected neurons. Neurons transmit signals to other neurons only when the intensity of the received signal is above a certain threshold.

Artificial neurons simulate the roles of these actual neurons through mathematical models. Each artificial neuron receives multiple signals, $x_1, x_2, \ldots, x_j$, composed of zeroes and ones, and it calculates the weighted sum of the signals that it receives according to their weights, $w_1, w_2, \ldots, w_j$. Depending on the model, signals belonging to the set $(-\infty, \infty)$ or the set $(0, \infty)$ may be received. A signal is then transmitted to the next artificial neuron only when the weighted sum of the signals received is above a certain intensity or threshold. The weights and thresholds of each neuron are determined by

the combination that leads to the best results based on past experience or data. An artificial neuron can be expressed as follows:

$$y_i = output = \begin{cases} 0 & if \ \sum_j w_j x_j \leq threshold \\ 1 & if \ \sum_j w_j x_j > threshold. \end{cases} \quad (10)$$

An artificial neural network is a machine learning methodology that can solve complex problems using a combination of simple artificial neurons. The network is composed of connections among multiple layers of artificial neurons, and each layer is divided into an input layer and an output layer, and a hidden layer between them. The process of training or optimizing the network involves determining the weights and thresholds for each artificial neuron to obtain the best results and therefore requires strong computational power.

Among the various possible structures of artificial neural networks, a model with several hidden layers is called a deep neural network. The term "deep learning" refers to the use of such deep neural networks for machine learning. This approach is intended to solve the problem of local minima that arises in the existing artificial neural network methodology. It is characterized by continuously using learnings from the data to improve the problem-solving ability of the network.

Default predictions using artificial neural networks are attempted before the 1990s. Yang et al. [44] explore several algorithms and show that Fisher discriminant analysis and probabilistic neural networks have the best prediction performance. Recently, following the pioneering work of Hinton, Osindero, and Teh [45], artificial neural networks are again emerging as a technique for corporate default predictions. Falavigna [46] predicts the default risks of small Italian companies with insufficient account information using an artificial neural network algorithm. López Iturriaga and Sanz [47] estimate and visualize banks' default risks by combining multilayer perceptrons and self-organizing maps. Azayite and Achchab [48] improve a neural network's default prediction model by incorporating discriminant variables. Geng et al. [49] show that the neural network approach has a better performance than other classifier algorithms. In addition, many studies have attempted to improve the prediction performance by revising the neural network algorithm [50–52].

In recent years, deep learning has used a convolutional neural network (CNN) and recurrent neural network (RNN) algorithms to solve the overfitting problems that arise in the learning process and improve performance. CNN algorithms divide the information used for learning into multiple domains and analyze the highly relevant domains using limited information. Thus, these algorithms perform well in the field of image recognition and are widely used. The RNN algorithms are a way to process and analyze data sequentially rather than independently analyzing the data used for learning, and they are used to predict time series data and recognize the context of text data. Currently, deep learning is used to understand the context of information and perform various tasks, such as pattern recognition, natural language processing, and autonomous driving.

3.4. Other Studies

Because a corporate default is a rare event, the training datasets are typically highly imbalanced. Quantitative models that use the full sample may not be appropriate because they may result in biased predictions [53,54]. To overcome this problem, Zhou [55] compares several sampling techniques (random oversampling with replication, the synthetic minority oversampling technique, random undersampling, and undersampling based on clustering from the nearest neighbor) across several machine learning algorithms (i.e., discriminant analysis, logistic regression, artificial neural networks, and SVM). The results show that random undersampling and SVM perform well in most cases, but the number of defaults in the training dataset can impact which methodology is most effective. Similarly, Veganzones and Séverina [56] also show that an imbalanced dataset can disturb prediction performance and that the SVM method is less sensitive than other methodologies. Kim et al. [57] suggest using the optimization of cluster-based undersampling to solve the imbalance problem. Piri et al. [58]

use a synthetic informative minority oversampling algorithm to enhance SVM performance with an imbalanced dataset. Tian et al. [59] claim that different sampling techniques are required depending on the purpose of the study. Song and Peng [60] suggest a multi-criteria decision making-based approach.

4. Discussion

In this study, we investigate previous studies on corporate default prediction. We also categorize them from the perspectives of financial engineering and machine learning. The findings are as follows. First, in much of the machine learning literature, corporate default predictions are considered as classification problems. On the other hand, structural and hazard models in the field of financial engineering understand and analyze corporate defaults as sequences. These approaches help us to understand how the time-varying changes in the variables affect a company's default risk. Second, macroeconomic factors are rare or often neglected in corporate bankruptcy prediction studies using machine learning methodologies. Of course, corporate defaults are mainly affected by the financial conditions of individual companies. However, macroeconomic conditions also have an important effect. Corporate default predictions using classification methodologies require the assumption that individual events are independent. However, if macroeconomic conditions change, such as the financial crisis, corporate defaults are no longer independent events. Third, in many cases, the machine learning approaches show superior corporate default forecasting performances to the financial engineering approaches. However, machine learning methodologies do not provide a meaningful answer to the cause of corporate default. While these forecasting can be useful for investors or credit rating agencies, they do not help executives who want to improve their businesses.

Our findings also suggest some new tasks in the field of machine learning when predicting corporate defaults. First, it is important to keep in mind that a corporate default prediction model is a multi-period model in which the future is affected by past decisions. Nevertheless, many machine learning models use one-period financial statements to avoid complexity. However, corporate defaults are not generally caused by one excessive loss or huge debt, as inadequate decision making over several periods often results in corporate default. Because the context is important for corporate default predictions, multi-period models are more appropriate for explaining corporate defaults; the RNN methodology can be a good example. In addition, it is necessary to use financial statement items from multiple periods at the same time. Second, the stock price and the corporate value evaluated in the stock market are important factors. However, the daily updated stock price is rarely used in machine learning approaches because that data cycle is inconsistent with that of financial statements. Even when the stock price is used, only the price at the time of the financial statements is incorporated. However, many financial research studies confirm that a company's stock price is an important explanatory variable for predicting corporate default. Finally, a corporate default prediction model should be able to suggest the cause of default. Predicting a corporate default is not simply a cats-or-dogs classification problem. Corporate executives and government officials can obtain a large amount of information from corporate default prediction models. However, if the model is a black box and cannot identify the issues leading to a default prediction, the model's usefulness is limited. Corporate default forecasts should not stop with classifications but rather should be able to offer clues on how to avoid defaults.

This study has several limitations. We do not cover all topics related to corporate default predictions. Explanatory variables that can affect corporate default are not considered. Variable selection techniques that can improve forecasting performances are also not covered. While we classified the related literature into statistical and machine learning approaches, we do not cover the entirety of the literature but present representative studies that are important in each category. Through this concentration, we derive meaningful findings and insights associated with this topic.

5. Conclusions

This review paper investigates the progress of research related to corporate default predictions and examines the main research methodologies by classifying them as either statistical models or

machine learning algorithms. In addition, it forecasts the financial sector innovation brought about by the development of machine learning. In doing so, this study aims to provide clues regarding the future convergence of research in the management science and computer science fields and to lay the foundation for expanding financial engineering methodologies to predict corporate default risks.

Our findings and suggestions for corporate default predictions are even more meaningful at this particular time. Owing to a series of technological developments referred to as the Fourth Industrial Revolution, the need to apply new methodologies to the field of financial engineering, including corporate default forecasting, is emerging. In particular, the big data analysis methodologies presented in this study suggest the need for enterprise-wide data governance to diagnose business conditions and help investors make accurate decisions. Not just large companies but also small and medium-sized enterprises now need to lay the foundation to easily and rapidly introduce new technologies by accurately identifying the type, size, and frequency of management data and continuously implementing quality management. The increase in the value of utilizing big data associated with the development of machine learning methodologies is accompanied by the need to strengthen the protection of personal information. Although de-identification measures are taken to protect personal information in the process of using big data, personal information has been leaked, in some cases through re-identification by combining big data with other information disclosed on the Internet. It is necessary to actively utilize the latest technologies, such as homogeneous cryptography, to reduce the risk of information leakage and perform big data analysis and processing without disclosing personal information.

Corporate default predictions using machine learning demand attention because the calculation process used to generate predictions may be a black box depending on the algorithm. Thus, although such methodologies can be used to calculate corporate default risk, they face the limitation that they cannot provide strategies for improving a company's management to reduce default risk. Thus, when performing corporate default forecasting, it is necessary to select an appropriate methodology that can provide suitable information for the purpose of prediction, requiring a detailed understanding of the appropriate utilization of each methodology.

Author Contributions: Proposal & original idea, H.K. and H.C.; conceptualization, H.K. and D.R.; modeling, H.K. and D.R.; methodology, H.K. and H.C.; validation, D.R.; resources, D.R.; software, H.K.; literature review, H.K., H.C. and D.R.; economic & business implication, D.R.; writing—original draft preparation, H.K., H.C., and D.R.; writing—review & editing, D.R.; discussion, H.K. and D.R.; project administration, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT; Ministry of Science and ICT) [No. 2019R1G1A1100196].

Acknowledgments: We are grateful for the valuable comments of three anonymous referees.

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

References

1. Beaver, W.H. Financial ratios as predictors of failure. *J. Account. Res.* **1966**, *4*, 71–111. [CrossRef]
2. Merton, R.C. On the pricing of corporate debt: The risk structure of interest rates. *J. Financ.* **1974**, *29*, 449–470.
3. Jessen, C.; Lando, D. Robustness of distance-to-default. *J. Bank. Financ.* **2015**, *50*, 493–505. [CrossRef]
4. Glover, B. The expected cost of default. *J. Financ. Econ.* **2016**, *119*, 284–299. [CrossRef]
5. Brogaard, J.; Li, D.; Xia, Y. Stock liquidity and default risk. *J. Financ. Econ.* **2017**, *124*, 486–502. [CrossRef]
6. Hillegeist, S.A.; Keating, E.K.; Cram, D.P.; Lundstedt, K.G. Assessing the probability of bankruptcy. *Rev. Account. Stud.* **2004**, *9*, 5–34. [CrossRef]
7. Altman, E.I. Financial Ratios, Discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [CrossRef]
8. Ohlson, J.A. Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res.* **1980**, *18*, 109–131. [CrossRef]

9. Duffie, D.; Saita, L.; Wang, K. Multi-period corporate default prediction with stochastic covariates. *J. Financ. Econ.* **2007**, *83*, 635–665. [CrossRef]
10. Alaka, H.A.; Oyedele, L.O.; Owolabi, H.A.; Kumar, V.; Ajayi, S.O.; Akinade, O.O.; Bilal, M. Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Syst. Appl.* **2018**, *94*, 164–184. [CrossRef]
11. Balcaen, S.; Ooghe, H. 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. *Br. Account. Rev.* **2006**, *38*, 63–93. [CrossRef]
12. Devi, S.S.; Radhika, Y. A survey on machine learning and statistical techniques in bankruptcy prediction. *Int. J. Mach. Learn. Comput.* **2018**, *8*, 133–139. [CrossRef]
13. Ravi Kumar, P.; Ravi, V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques—A review. *Eur. J. Oper. Res.* **2007**, *180*, 1–28. [CrossRef]
14. Altman, E.I. *Corporate Financial Distress and Bankruptcy: A Complete Guide to Predicting and Avoiding Distress and Profiting from Bankruptcy*, 2nd ed.; John Wiley and Sons Inc.: New York, NY, USA, 1993.
15. Mare, D.S.; Moreira, F.; Rossi, R. Nonstationary Z-score measures. *Eur. J. Oper. Res.* **2017**, *260*, 348–358. [CrossRef]
16. Charitou, A.; Neophytou, E.; Charalambous, C. Predicting corporate failure: Empirical evidence for the UK. *Eur. Account. Rev.* **2004**, *13*, 465–497. [CrossRef]
17. Zmijewski, M.E. Methodological issues related to the estimation of financial distress prediction models. *J. Account. Res.* **1984**, *22*, 59–82. [CrossRef]
18. Foreman, R.D. A logistic analysis of bankruptcy within the US local telecommunications industry. *J. Econ. Bus.* **2003**, *55*, 135–166. [CrossRef]
19. Campbell, J.Y.; Hilscher, J.; Szilagyi, J. In search of distress risk. *J. Financ.* **2008**, *63*, 2899–2939. [CrossRef]
20. Aretz, K.; Florackis, C.; Kostakis, A. Do stock returns really decrease with default risk? New International Evidence. *Manag. Sci.* **2018**, *64*, 3821–3842. [CrossRef]
21. Kukuk, M.; Rönnberg, M. Corporate credit default models: A mixed logit approach. *Rev. Quant. Financ. Account.* **2013**, *40*, 467–483. [CrossRef]
22. Bonfim, D. Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *J. Bank. Financ.* **2009**, *33*, 281–299. [CrossRef]
23. Shumway, T. Forecasting bankruptcy more accurately: A simple hazard model. *J. Bus.* **2001**, *74*, 101–124. [CrossRef]
24. Cox, D.R. Regression models and life tables (with discussion). *J. R. Stat. Soc.* **1972**, *34*, 187–220.
25. Chava, S.; Jarrow, R.A. Bankruptcy prediction with industry effects. *Rev. Financ.* **2004**, *8*, 537–569. [CrossRef]
26. Nam, C.; Kim, T.; Park, N.; Lee, H. Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *J. Forecast.* **2008**, *27*, 493–506. [CrossRef]
27. Dakovic, R.; Czado, C.; Berg, D. Bankruptcy prediction in Norway: A comparison study. *Appl. Econ. Lett.* **2010**, *17*, 1739–1746. [CrossRef]
28. Duan, J.; Sun, J.; Wang, T. Multiperiod corporate default prediction—A forward intensity approach. *J. Econom.* **2012**, *170*, 191–209. [CrossRef]
29. Traczynski, J. Firm default prediction: A Bayesian model-averaging approach. *J. Financ. Quant. Anal.* **2017**, *52*, 1211–1245. [CrossRef]
30. Figlewski, S.; Frydman, H.; Liang, W. Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *Int. Rev. Econ. Financ.* **2012**, *21*, 87–105. [CrossRef]
31. Tian, S.; Yu, Y.; Guo, H. Variable selection and corporate bankruptcy forecasts. *J. Bank. Financ.* **2015**, *52*, 89–100. [CrossRef]
32. Samuel, A.L. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* **1959**, *3*, 210–229. [CrossRef]
33. Mitchell, T. *Machine Learning*; McGraw Hill: New York, NY, USA, 1997.
34. Barboza, F.; Kimura, H.; Altman, E. Machine learning models and bankruptcy prediction. *Expert Syst. Appl.* **2017**, *83*, 405–417. [CrossRef]
35. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning, 112*; Springer: New York, NY, USA, 2013.
36. Shin, K.S.; Lee, T.S.; Kim, H.J. An application of support vector machines in bankruptcy prediction model. *Expert Syst. Appl.* **2005**, *28*, 127–135. [CrossRef]

37. Chen, M.Y. Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Comput. Math. Appl.* **2011**, *62*, 4514–4524. [CrossRef]
38. Liang, D.; Lu, C.C.; Tsai, C.F.; Shih, G.A. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *Eur. J. Oper. Res.* **2016**, *252*, 561–572. [CrossRef]
39. Lu, Y.; Zhu, J.; Zhang, N.; Shao, Q. A hybrid switching PSO algorithm and support vector machines for bankruptcy prediction. In Proceedings of the 2014 International Conference on Mechatronics and Control, Jinzhou, China, 3–5 July 2014; pp. 1329–1333.
40. Olson, D.L.; Delen, D.; Meng, Y. Comparative analysis of data mining methods for bankruptcy prediction. *Decis. Support Syst.* **2012**, *52*, 464–473. [CrossRef]
41. Tsai, C.F.; Hsu, Y.F.; Yen, D.C. A comparative study of classifier ensembles for bankruptcy prediction. *Appl. Soft Comput.* **2014**, *24*, 977–984. [CrossRef]
42. Zięba, M.; Tomczak, S.K.; Tomczak, J.M. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Syst. Appl.* **2016**, *58*, 93–101. [CrossRef]
43. Du Jardin, P. Failure pattern-based ensembles applied to bankruptcy forecasting. *Decis. Support Syst.* **2018**, *107*, 64–77. [CrossRef]
44. Yang, Z.; Platt, M.B.; Platt, H.D. Probabilistic neural networks in bankruptcy prediction. *J. Bus. Res.* **1999**, *44*, 67–74. [CrossRef]
45. Hinton, G.E.; Osindero, S.; Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
46. Falavigna, G. Financial ratings with scarce information: A neural network approach. *Expert Syst. Appl.* **2012**, *39*, 1784–1792. [CrossRef]
47. López Iturriaga, F.J.; Sanz, I.P. Bankruptcy visualization and prediction using neural networks: A study of US commercial banks. *Expert Syst. Appl.* **2015**, *42*, 2857–2869. [CrossRef]
48. Azayite, F.Z.; Achchab, S. Hybrid discriminant neural networks for bankruptcy prediction and risk scoring. *Procedia Comput. Sci.* **2016**, *83*, 670–674. [CrossRef]
49. Geng, R.; Bose, I.; Chen, X. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *Eur. J. Oper. Res.* **2015**, *241*, 236–247. [CrossRef]
50. Chen, J.-H. Developing SFNN models to predict financial distress of construction companies. *Expert Syst. Appl.* **2012**, *39*, 823–827. [CrossRef]
51. Chen, H.-J.; Huang, S.Y.; Lin, C.-S. Alternative diagnosis of corporate bankruptcy: A neuro fuzzy approach. *Expert Syst. Appl.* **2009**, *36*, 7710–7720. [CrossRef]
52. Chen, N.; Ribeiro, B.; Vieira, A.S.; Duartea, J.; Neves, J.C. A genetic algorithm-based approach to cost-sensitive bankruptcy prediction. *Expert Syst. Appl.* **2011**, *38*, 12939–12945. [CrossRef]
53. Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *30*, 25–36.
54. Lane, P.C.R.; Clarke, D.; Hender, P. On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decis. Support Syst.* **2012**, *53*, 712–718. [CrossRef]
55. Zhou, L. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowl. Based Syst.* **2013**, *41*, 16–25. [CrossRef]
56. Veganzones, D.; Séverina, E. An investigation of bankruptcy prediction in imbalanced datasets. *Decis. Support Syst.* **2018**, *112*, 111–124. [CrossRef]
57. Kim, H.-J.; Jo, N.-O.; Shin, K.-S. Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction. *Expert Syst. Appl.* **2016**, *59*, 226–234. [CrossRef]
58. Piri, S.; Delen, D.; Liu, T. A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decis. Support Syst.* **2018**, *106*, 15–29. [CrossRef]
59. Tian, S.; Yu, Y.; Zhou, M. Data sample selection issues for bankruptcy prediction. *Risk Hazards Crisis Pub. Policy* **2015**, *6*, 91–116. [CrossRef]
60. Song, Y.; Peng, Y. A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction. *IEEE Access* **2019**, *7*, 84897–84906. [CrossRef]