Corporate Bankruptcy Predictions Using CEO Social Network Information

Tsung-Kang Chen[*]
E-mail: vocterchen@nycu.edu.tw
Department of Management Science, National Yang Ming Chiao Tung University
Center for Research in Econometric Theory and Applications, National Taiwan University

Ting-Ru Chang
E-mail: ctr99.mg10@nycu.edu.tw
Department of Management Science, National Yang Ming Chiao Tung University

Yun Hao
E-mail: peterhao.c@nycu.edu.tw
Department of Management Science, National Yang Ming Chiao Tung University

Yu-Chun Lin
Email: magic.mg09@nycu.edu.tw
Department of Management Science, National Yang Ming Chiao Tung University

[*] Corresponding author.

Corporate Bankruptcy Predictions Using CEO Social Network Information

Abstract

Different from previous bankruptcy prediction literature, this study introduces CEO's and other executives' social network characteristics information (El-Khatib et al., 2015) in addition to Barboza's et al. (2017) 11 financial variables to investigate whether CEO's and other executives' social network characteristics information improve bankruptcy prediction effectiveness. This study implements the bankruptcy prediction analyses using machine learning models and U.S. firm observations from 2000 to 2020. Empirical results of this study show that each machine learning model significantly improves the effectiveness of short-term bankruptcy prediction after introducing CEO's and other executives' social network characteristics according to the performance measures of F1-score and AUC. This is mainly because the degree of the CEO's controls over information flow (namely the degree of manipulating the speed of information flow transmission, which thus shapes incomplete information) has the highest information content for corporate bankruptcy among the social network characteristic variables. Hence, the finding that CEO social network characteristic variables can significantly improve the effectiveness of short-term bankruptcy prediction is consistent with Duffie and Lando (2001). Finally, the findings are still robust when considering different random states.

Keywords：CEO social network characteristics, Bankruptcy prediction, Machine learning, Network centrality, Incomplete information

1. Introduction

Since the subprime financial crisis in 2008, the adverse impact of corporate bankruptcy events on the economic system has become a significant concern in both academic and practical spheres. Among these concerns, particular attention has been given to the effective prediction of corporate bankruptcy events in advance. This is because the ability to predict corporate bankruptcy risk can not only assist financial institutions in making more accurate lending decisions, thereby reducing economic losses resulting from corporate bankruptcies, but also help investment institutions identify potential risks associated with target companies in advance, thereby conducting correct investment and risk management decisions. Therefore, the effectiveness of bankruptcy prediction can enhance the operating efficiency of banks in their credit loan business.

The current bankruptcy prediction models include both traditional statistical models (e.g., Beaver, 1966; Altman, 1968; Ohlson, 1980) and machine learning models (e.g., Barboza et al., 2017), with the latter being the predominant trend in development. Numerous scholars are currently dedicated to incorporating financial and non-financial information into corporate bankruptcy predictions with machine learning models. The financial and non-financial information include financial variables (Barboza et al., 2017), corporate governance variables (e.g., Liang et al., 2017), word frequency in annual reports' MD&A (management discussion and analysis) sections (e.g., Mai et al., 2019; Kim and Yoon, 2021), and text-based communication value (hereafter denoted as TCV) of annual reports (e.g., Chen et al., 2023a), and TCV uncertainty of annual reports (e.g., Chen et al., 2023b). However, there is currently a scarcity of studies incorporating CEO characteristics into bankruptcy prediction models, especially from social capital perspective. As the primary decision-maker in a company, the CEO's traits significantly impact the operating performance and risks of the firm business. Among these traits, the CEO's social network can provide the necessary resources for managerial decision-making, subsequently influencing firm asset value distributions (e.g., Bloch et al., 2008; Ferris et al., 2019), debt financing decisions, information transparency (e.g., Ferris et al., 2017), and credit risk (e.g., Chen & Tseng, 2022). According to structural-form credit risk models (Merton, 1974; Duffie and Lando, 2001), the main determinants of firm credit risk are asset value, asset value volatility, leverage ratio, and incomplete accounting information. Therefore, it is reasonably anticipated that CEO social network characteristics contribute to improving the predictive power of corporate bankruptcy. The main research purpose of this study is to employ the 11 financial variables proposed by Barboza et al. (2017) as a benchmark model and further investigate whether the inclusion of CEO social network characteristics variables can effectively enhance the predictive power of corporate bankruptcy.

Social capital is a form of social resource embedded in relationships, and individuals can acquire these resources through social networks (Anderson et al., 2007). In the social sciences, network centrality has been considered a source of influence and power. In the corporate context, the network characteristics established by CEOs, whether through educational background, work experience, or relationships formed through social activities, may influence corporate decision-making. Taking the example of social networks between CEOs and board members. When a firm performs poorly, social networks between the CEOs and board members may reduce the board's motivation to decrease CEOs' compensations or terminate the CEOs, which solidify the CEOs' position, enhance their power, and thus stimulate the CEOs to adopt riskier strategies (Fan et al., 2021). Moreover, CEOs with high social network centrality, due to their greater power, can influence firm value and increase risk by controlling information flow, potentially depriving other shareholders' rights for their private benefits (e.g., agency theory) (El-Khatib et al., 2015). Based on the above discussions, it is evident that CEO social network characteristics (e.g., centrality) can affect financial decision-making, operating risk and credit risk, consequently impacting the interests of creditors and increasing the likelihood of corporate bankruptcy.

Different from existing machine learning-based bankruptcy prediction literature that utilizes non-financial information such as textual features from annual reports (e.g., Mai et al., 2019; Kim and Yoon, 2021; Chen et al., 2023a; Chen et al., 2023b) or corporate governance information (Liang et al., 2017), this study explores whether including CEO social network characteristics enhance the effectiveness of bankruptcy prediction. In this regard, this study follows Zhang et al. (2023) in using network centrality as the main measure of CEO social network characteristics and adopts the metrics of Degree, Betweenness, Closeness, and Eigenvector based on El-Khatib et al. (2015) to quantify network centrality. The four network centrality measures are defined as follows: Degree represents the number of direct connections an individual has with other individuals; Betweenness reflects the degree to which an individual controls the flow of information; Closeness signifies the efficiency with which an individual can obtain information from others; and Eigenvector represents the importance of an individual in the network. Therefore, CEO social network centrality variables not only influence a firm’s asset value and risk (e.g., Degree, Eigenvector, Closeness) but also reflect the degree of information asymmetry within the firm (e.g., Betweenness). Hence, based on the theoretical framework of structural-form credit risk models (e.g. Merton, 1974; Duffie and Lando, 2001), this study hypothesizes that CEO social network centrality variables enhance the effectiveness of corporate bankruptcy prediction. In addition, this study also considers social network characteristics of other managers and further analyzes which managers’ social network centrality variables are most significant for predicting corporate bankruptcy.

This study employs U.S. 31,525 firm observations data from year 2000 to 2020 (including 31,161 non-bankrupt firms and 364 bankrupt firms) to investigate whether CEO social network centrality variables contribute to the improvement of bankruptcy prediction effectiveness. This study also considers the issue of corporate bankruptcy term structure and examines the comparative analysis of the predictive effectiveness of CEO social network centrality variables on short-term and long-term corporate bankruptcy. Specifically, this study uses the period from 2000 to 2014 as the training period (21,373 non-bankrupt firms and 328 bankrupt firms) and the period from 2015 to 2020 as the testing period (9,788 non-bankrupt firms and 36 bankrupt firms). The training period employs a rolling approach on an annual basis to predict corporate bankruptcy in future periods. Furthermore, this study considers 30 sets of random states to test whether including CEO social network centrality variables significantly enhances the effectiveness of corporate bankruptcy prediction. In terms of model performance measurement, this research adopts F1-score and AUC (Area under curve) as main indicators for assessing predictive effectiveness. The empirical results indicate that, compared to the bankruptcy prediction model using the 11 financial variables proposed by Barboza et al. (2017), the inclusion of CEO and other managerial social network characteristics significantly improves the F1-score of the model. For instance, in predicting whether a firm will go bankrupt in the next year, adding CEO and other managerial social network characteristics enhances the predictive performance of both Random Forest (RF) and Extreme Gradient Boosting (XGBoost) machine learning models with various data imbalance processing methods. Especially when XGBoost model is combined with the RandomOverSampler method, the model shows the highest prediction performance (e.g., its F1-score increases from 0.0819 to 0.3752).

In terms of AUC, the RF or XGBoost models combined with the EasyEnsemble method exhibit the highest prediction performance (with AUC values of 0.8712 and

0.8688, respectively), demonstrating a significant increase compared to the model setting that utilizing financial variables as model input variables. In addition, this study also finds that, under the model setting of XGBoost with EasyEnsemble method, the number of Type I errors (namely predicting non-bankruptcy, but actually bankruptcy) slightly increases from 0.3056 to 1.0778 whereas the number of Type II errors (namely predicting bankruptcy, but actually non-bankruptcy) significantly decreases from 482.0945 to 149.9444. These results indicate that CEO and other managerial social network characteristics can substantially reduce the number of Type II errors while maintaining a negligible number of Type I errors. Furthermore, this study reveals that the social network characteristics of CEOs and other managers demonstrate a larger improvement on the short-term bankruptcy predictions than long-term bankruptcy predictions. In long-term bankruptcy predictions, although the number of Type II errors still experiences a substantial reduction, the improvement in the number of Type I errors diminishes. Therefore, the findings suggest that the social network characteristics of CEOs and other managers can help financial institutions improve the efficiency of fund utilization in corporate loan business and make accurate credit loan decisions.

Moreover, to gain further insights into which social network characteristics of CEOs and other managers are crucial bankruptcy prediction indictors, this study conducts feature engineering on the introduced research variables. The empirical results reveal that the CEO's control over information flow (namely CEO Betweenness) and the standard deviation of the social network size among non-CEO managers (namely standard deviation of non-CEO managers' Degree) are the top two influential variables among the social network characteristics variables. Besides, these two features rank as the fourth to sixth most important when considering all variables, including the 11 financial variables proposed by Barboza et al. (2017). Given that CEO Betweenness signifies the degree to which the CEO controls information flow, it can represent the

CEO's manipulation of the speed of information transmission, impacting the communicative value of information received by external investors and thus influencing firm performance under incomplete information. On the other hand, the standard deviation of the social network size among other managers affects the variance in external investors' assessments of the firm value distribution. For the variable of CEO Betweenness, the most crucial bankruptcy prediction variable among the social network characteristics variables, it reflects the communicative value of corporate information and thus can be used to describe the degree of incomplete information about the firm. Hence, in line with the theoretical perspectives of Duffie and Lando's (2001) incomplete information-based structural form credit risk model, it can be inferred that CEO social network characteristics variables have a more noticeable improvement effect on short-term bankruptcy prediction effectiveness. The empirical results also indicate that the social network characteristics of CEOs and other managers exhibit a better prediction performance for short-term corporate bankruptcy than long-term corporate bankruptcy, consistent with the theoretical perspectives of Duffie and Lando (2001) and Yu (2005).

The potential contributions of this study include: (1) introducing CEO and other managers' social network characteristics variables and incorporating the concepts of structural credit risk models (Merton, 1974; Duffie and Lando, 2001) into machine learning-based bankruptcy prediction models; (2) elaborating on the economic implications of CEO and other managers' social network centrality variables in firm credit risk, especially the CEO Betweenness variable; (3) providing the empirical evidence that CEO and other managers' social network characteristics variables significantly enhance bankruptcy prediction effectiveness, particularly in short-term bankruptcy prediction scenarios; (4) demonstrating that CEO and other managers' social network characteristics variables significantly improve the number of misjudging non-bankrupt firms as bankrupt ones (Type II error); (5) offering the evidence of the ranking of the importance of CEO and other managers' social network characteristics variables in bankruptcy prediction, with CEO Betweenness variable (degree of CEO's control over information flow) being of the highest importance. Given that the CEO Betweenness variable represents the CEO's ability to control the speed of information transmission (i.e., communicative value), it affects the degree of corporate incomplete information. The finding that CEO social network characteristics variables significantly improve short-term bankruptcy prediction effectiveness aligns with Duffie and Lando (2001) that the impact of incomplete information on short-term credit risk assessment is more significant. Therefore, this study provides theoretical foundations and economic implications for understanding the impact of CEO and other managers' social network characteristics variables on the prediction of bankruptcy term structure.

2. Literature Review

This study aims to explore whether incorporating CEO and other managers' social network characteristics variables in machine learning-based bankruptcy prediction models can improve the model effectiveness of corporate bankruptcies predictions. This section introduces the relevant literature from three aspects, including (1) the related researches on corporate credit/bankruptcy risk, (2) the researches on the relationship between CEO social capital and corporate credit/bankruptcy risk, and (3) the researches on corporate bankruptcy predictions using machine learning modes. The details are demonstrated as follows.

2.1. Researches on Corporate Credit/Bankruptcy Risk

2.1.1. Financial Variables in Annual Reports and Corporate Credit/Bankruptcy Risk

To determine whether a firm is facing bankruptcy, one of the most direct approaches is to analyze the firm's financial health. Beaver (1966) proposes a univariate analysis using a dichotomous classification test and employing 30 financial variables to predict corporate bankruptcy across 158 sample firms. Beaver (1966) finds that "cash flow/total liabilities" has the best predictive power in corporate bankruptcy prediction. However, the accuracy of predicting bankrupt and non-bankrupt companies using financial ratios is not stable, leading to potential contradictions. Hence, subsequent literature largely integrates multiple financial variables for researches. For instance, Altman (1968) utilizes multiple discriminant analysis and introduces the Z-score model, encompassing five crucial financial variables of “Net Working Capital to Total Assets ratio” (NWC_TA), “Retained Earnings to Total Assets ratio” (RE_TA), “Earnings Before Interest and Taxes to Total Assets ratio” (EBIT_TA), “Equity Market Value to Total Debt ratio” (EMV_Debt), and “Sales to Total Assets ratio” (Sales_TA). The Z-score model serves as a tool to assess the severity of corporate bankruptcy but does not provide the probability of bankruptcy. Subsequently, Ohlson (1980) utilizes logistic regression model and introduces corporate financial variables to predict corporate bankruptcy, which provides bankruptcy probability and scores (O-score). However, traditional statistical models are limited by theoretical probability distributions, which hinders further improvements in predictive power in practical applications.

Barboza et al. (2017), based on machine learning models, utilize Altman's (1968) Z-score with its five financial variables and an additional six financial variables from Carton and Hofer (2006) to conduct a comparative analysis of bankruptcy prediction effectiveness using U.S. firm observations from 1985 to 2013. The results reveal that their predictive effectiveness exceeds those of traditional statistical methods.

2.1.2. Annual Report Textual Characteristics and Corporate Credit/Bankruptcy Risk

In the credit risk literature, aside from financial information in annual reports, non-financial information in annual reports, especially textual statements, has also received considerable attention. Since financial statements are a crucial means for information users and investors to gain insights into a firm's operation status and development trends, the textual explanations that clarify the implications of financial statement numbers have corresponding communicative value for external investors (Seebeck and Kaya, 2023). Among the variables used to measure the communicative value of annual reports, readability and evaluative content are frequently employed in past literature, as seen in studies such as Bonsall and Miller (2017), Ertugrul et al. (2017), and Chen and Tseng (2021). Bonsall and Miller (2017) find that lower annual report readability leads to poorer credit rating scores (higher default risk), greater divergence in opinions among bond rating agencies, and higher costs of debts. Ertugrul et al. (2017) discover that lower readability or greater ambiguous tones in annual reports results in higher borrowing costs. Chen and Tseng (2021) also demonstrate that higher readability of notes to consolidated financial statements leads to lower corporate bond spreads. Therefore, based on the above discussions, it can be inferred that lower readability or greater ambiguous tones of financial information disclosures in annual reports may increase the probability of corporate default.

Subsequently, in research applying machine learning models to bankruptcy prediction, Mai et al. (2019) and Kim and Yoon (2021) introduce the word frequency of the Management Discussion and Analysis (MD&A) section in annual reports. Chen et al. (2023a) and Chen et al. (2023b) incorporate the text-based communicative value (hereafter denoted as TCV) of annual reports and its uncertainty (namely annual report TCV uncertainty) as new input variables into bankruptcy prediction models and find that both TCV level and TCV uncertainty improve the effectiveness of corporate bankruptcy prediction models. The above studies all confirm that textual information in annual reports indeed enhances the effectiveness of bankruptcy predictions.

In addition to the annual report textual information, managers' characteristics also significantly influence a firm’s business strategies and the related risks. For instance, CEO overconfidence has an impact on financing decisions (e.g., Malmendier et al., 2011; Lin et al., 2020) and investment decisions (e.g., Malmendier and Tate, 2005). CEO social capital (e.g. networks) also influences a firm’s risky activities and policies (Ferris et al., 2017; Ferris et al., 2019). CEO social networks can provide the resources needed for managerial decisions and thus further affect a firm’s asset value distribution (e.g., Bloch et al., 2008; Ferris et al., 2019), debt financing decisions, information transparency (e.g., Ferris et al., 2017), and credit risk (e.g., Chen & Tseng, 2022). This study therefore uses the 11 financial variables proposed by Barboza et al. (2017) as a benchmark model and further investigates whether the social networks of CEOs and other managers contribute to enhancing bankruptcy prediction effectiveness.

2.2. Research on Social Capital and Social Network

2.2.1. Social Capital

Social capital is multidimensional, composed of various social factors, including networks, trust, norms, etc. In addition, social capital can be viewed as a resource embedded in social structures, obtained through purposeful actions. Burt (2001) demonstrates that the assets an individual acquires due to their specific positions in social networks constitute social capital. Nahapiet and Ghoshal (1998) categorize social capital into three dimensions: cognitive, relational, and structural. Moreover, social capital is also associated with executive compensations, risk-taking behaviors, financing policies, corporate innovation capabilities, and corporate strategies (Nahapiet and Ghoshal, 1998; Xing et al., 2023).

Social capital can be divided into internal social capital and external social capital. The former exists within a group that shares common interests or goals, and the latter exists between two different groups. Internal social capital is believed to enhance the teamwork and collaboration among board members within the board. External social capital may exist between two firms, assisting the firm in gaining novel strategies (Xing et al., 2023). However, in specific situations, social capital may be useless or even harmful (Coleman, 1988).

2.2.2. Social Network

According to the definition of Ferris et al. (2017), social network is a medium for creating, maintaining, and utilizing social capital. Through social network, information can be obtained at a lower cost, helping to reduce information asymmetry. In addition, social capital also provides an informal insurance mechanism in social networks (Bloch et al., 2008), which may which may encourage CEOs who prefer high risks to make decisions that are unfavorable to the firm, thereby increasing the risks borne by the firm.

Network centrality is currently one of the concepts commonly discussed in social networks (Borgatti, 2005). It indicates the importance of a node in the network and can be measured by various indicators. The most common four indicators of network centrality are Degree, Betweenness, Closeness, and Eigenvector. Among these four indicators, Degree represents the number of direct connections an individual has with others; Betweenness indicates the extent to which an individual controls the flow of information; Closeness represents the efficiency with which an individual can obtain information from others; and Eigenvector signifies the importance of an individual in the network (El-Khatib et al., 2015). According to Zhang et al. (2023), this study defines CEO’s social capital as the relative position of the CEO in the social network, measured by network centrality.

2.2.3. The Relationship Between Managerial Social Networks and Credit Risk

Ferris et al. (2017) and Ferris et al. (2019) demonstrate that the broader a manager’s social network, the more information they have access to. Hence, managers may leverage this advantage to make better decisions or engage in higher-risk investment activities that may damage company value. In addition, Chen and Tseng (2022) develop a measure of social resources inequality among management team members, defined as the relative ratios of social capital among management team members. A higher value of the relative ratio implies more unevenly distributed social capital among management team members (i.e., greater inequality in power), which thus leads to an increase in corporate operating risks. Moreover, Chen and Tseng (2022) also find that (1) a larger social network of top management teams (TMT) members leads to more status conflicts, increasing corporate credit risk; (2) the power of CEOs intensifies the positive relationship between social networks and corporate credit risk; (3) a broader social network for middle managers (MM) enhances the information-sharing mechanism, thereby reducing corporate credit risk. In summary, the characteristics of CEOs' social networks, as well as the relative ratio or variance of social networks among management teams, impact a firm's credit risk.

2.3. Applying Machine Learning Models to Corporate Bankruptcy Predictions

The current application of machine learning models in bankruptcy prediction can be broadly categorized into two types: one involves using more complex machine learning or deep learning models to enhance predictive power on the basis of existing input variables, and the other introduces additional input variables to improve predictive effectiveness based on the existing machine learning models.

The latter not only enhances predictive power but also provides interpretability and economic significance, such as financial variables (Barboza et al., 2017), corporate governance variables (e.g., Liang et al., 2017), word frequency of the MD&A section in annual reports (e.g., Mai et al., 2019; Kim and Yoon, 2021), and TCV levels and TCV uncertainty of annual reports (e.g., Chen et al., 2023a; Chen et al., 2023b). Alaka et al.

(2018) collect literature from 2010 to 2015 involving 49 studies that apply machine learning models for corporate bankruptcy prediction. The results show that no single model significantly outperforms others. Meanwhile, most of previous studies focus on enhancing the accuracy of bankruptcy prediction models through model refinement, which makes the models increasingly complex. However, this method overlooks the practical orientation of bankruptcy prediction, which aims to provide reasonable explanations for the prediction results (Sun et al., 2014). Regulatory frameworks such as Basel II, Basel III, and regulatory authorities all require financial institutions to adopt models with higher interpretability rather than models with higher predictive power.

In summary, this study employs Barboza’s et al. (2017) 11 financial variables as the benchmark model setting of bankruptcy predictions and introduces additional information on the social network characteristics of CEOs and other managers as new model input variables. In addition, this study incorporates structural-form credit risk models (Merton, 1974; Duffie and Lando, 2001) as the theoretical foundation and provides the bankruptcy prediction model with new interpretative aspects and economic significance. Furthermore, this study explores the prediction effectiveness comparisons for the effects of including CEO and other managerial social network characteristics on bankruptcy term structure. This constitutes one of the main contributions of the present research.

3. Research Methodology

The research procedures of this study are primarily outlined in the following steps. First, data collection involves obtaining social network information of CEOs and other managers from the BoardEx database while financial variables data and the information for identifying bankrupt firms are both sourced from the COMPUSTAT database. In addition, the sample period spans from the year 2000 to 2020. Second, data pre-processing is conducted, including the removal of data with missing values, and the segmentation of sample data into the training set and the testing set on an annual basis. This study also addresses the problem of data imbalance in the training set data and uses a year-by-year rolling method to train the training data set. Finally, the study employs various indicators to assess the model's performance.

3.1. Data and variables

3.1.1. Dependent variable (Brupt)

We collect all U.S. bankrupt and non-bankrupt firm data from 2000 to 2020 as the preliminary research sample from Compustat database. We introduce the DLRSN CODE from Compustat database to classify the firms with DLRSN CODE 02 or 03 as bankrupt firms and the rest as non-bankrupt firms. Therefore, we define the Brupt variable as a dummy variable of bankrupt firms that labels bankrupt firms as 1 and non-bankrupt firms as 0. After data processing, our research dataset finally includes 364 bankrupt firms and 31,161 non-bankrupt firms.

3.1.2. Social network characteristics variables (SN_C)

According to literature such as El-Khatib et al. (2015) and Fan et al. (2021), it has been indicated that a firm’s executives’ social networks and network centrality both increase the firm’s risk. In addition, Chen & Tseng (2022) find a positive relationship between the relative ratios of social networks among different top management teams and corporate credit risk. This study follows El-Khatib et al. (2015), Fan et al. (2021), and Chen & Tseng (2022) to define: (1) CEO social network characteristics (SN_C_CEO), (2) other managers’ social network characteristics (SN_C_nCEO), (3) variance in other managers’ social network characteristics (std_SN_C_nCEO), (4) all managers’ social network characteristics (SN_C_all), and (5) variance in all managers’ social network characteristics (std_SN_C_all). The detailed definitions of each managers' social network characteristic variable are presented in Table 1. In total, there are 20 variables in these five categories. The social network characteristics are primarily measured using centrality indicators, including Degree, Betweenness, Closeness, and Eigenvector.

[Insert Table 1 in here]

3.1.3. Financial variables (FIN)

We follow Barboza et al. (2017) to introduce 11 financial variables as the input variables of the benchmark model setting. The 11 financial variables include five Altman's (1968) Z score component variables and six financial changes variables proposed by Carton and Hofer (2006). The five component variables of Altman's (1968) Z score include the ratio of net working capital to total assets (NWC_TA), the ratio of retained earnings to total assets (RE_TA), return on assets (EBIT_TA), the ratio of equity market value to total debts (EMV_Debt), and the ratio of net sales to total assets (Sales_TA). The six financial changes variables of Carton and Hofer (2006) cover the ratio of earnings before interests and taxes to net sales (namely EBIT margin, EBIT_Sales), the change rate of total assets (TA_growth), the change rate of net sales (Sales_growth), the change rate of the number of employee (EMP_growth), the change in return on equity (ROE_Chg), and the change in equity market-to-book value ratio (PB_Chg). The detailed definitions of the above 11 financial variables (Barboza et al., 2017) are shown in Table 2.

[Insert Table 2 in here]

3.2. Data pre-processing

The data pre-processing in this study includes handling missing values, splitting the dataset into the training and testing sets on an annual basis, and addressing data imbalance issues. Data pre-processing is carried out to ensure the integrity of the data, enabling it to fit the model and enhance the performance of the model. This section will provide a detailed description of each step in the data pre-processing procedures.

3.2.1. Null value processing

As this study primarily investigates whether the effectiveness of the bankruptcy prediction model is significantly improved by incorporating the social network characteristic variables of CEOs and other managers in addition to the existing benchmark model (Barboza et al., 2017), the social network characteristic variables of CEOs and other managers are considered essential variables. Therefore, this study removes samples with missing values for these types of social network characteristics variables. In addition, to ensure data completeness, samples with missing values for the 11 financial variables of Barboza et al. (2017) are also excluded. Finally, the data set contains 31,525 valid observations.

3.2.2. Data splitting

We follow Chen et al. (2023) to employ the time basis to split the research sample observations into the training data set and the testing data set for performing machine learning models. The training data set is used to develop the effective machine learning models of bankruptcy prediction while the testing data set is used to verify the forecast results of bankruptcy prediction models. We employ the sample observations during period from 2000 to 2014 as the baseline training data set and those during the rolling period year by year from 2015 to 2020 as testing data set. As mentioned in Chen et al. (2023), the reasons for employing time as sample data splitting basis include: (1) the information related to corporate financial structure are time-varying; and (2) the data properties of time series data are easier to be learned when using continuous time interval data to train the machine models.

We use the period from 2000 to 2014 as the baseline sample period of training data and the next year to the next six years as the sample periods of testing data. In addition, we adopt the method of year-by-year rolling adjustment to determine the sample periods of training data and testing data. For example, to perform corporate bankruptcy prediction in the next one year (namely 2015, 2016; 2017; 2018; 2019; 2020), we employ the year-by-year rolling baseline sample period (namely 1996 to 2014; 1996 to 2015; 1996 to 2016; 1996 to 2017; 1996 to 2018; 1996 to 2019). Moreover, we follow this similar methodology to perform bankruptcy predictions for the next two years, the next three years, the next four years, the next five years, and the next six years. The prediction effectiveness analysis of future periods is expressed through an average method.

3.2.3. Imbalanced data processing

Since the distribution of bankrupt and non-bankrupt firm data is generally imbalanced, imbalanced data processing becomes an important issue among bankruptcy prediction studies. For example, our research sample observations include 31,161 non-bankrupt firms and 364 bankrupt firms, implying that only 1.92% of the full sample firms are bankrupt and thus the data imbalance concern becomes a severe issue in this study. Therefore, we follow Chen et al. (2023) to employ EasyEnsemble, BalanceBaggingClassifier, RandomUnderSampling, RandomOverSampler, SMOTE, and SMOTEENN algorithms to mitigate the data imbalance concerns. Please see Table 3 for the sample distribution table of each year.

[Insert Table 3 in here]

The details of these six data imbalance methods are demonstrated in the following. Regarding the EasyEnsemble algorithm, it is an integrated learning algorithm composed of Bagging and Adaboost methods and its main concept is stated in order as follows: (1) performing random sampling k times in the majority category sample; (2) taking the same number of samples as the minority category sample in each sampling and generating k datasets; (3) training these data sets k times and generating k different models; (4) performing a majority vote to obtain the final result. In addition, we use EasyEnsemble algorithm to generate 1000 subsets by randomly sampling 1000 times, and then generates 1000 sub-models for majority decision to determine the final result.

Concerning the BalanceBaggingClassifier algorithm (Hido et al., 2009), it is a combination method of base classifier and EasyEnsemble algorithm. It has to be noted that the base classifier can be set employing the parameter setting base_estimator. Finally, we perform the final classification results by combining multiple models.

In addition, the concept of Random Undersampling primarily involves randomly removing samples from the majority class to achieve class balance by ensuring that all classes have the same quantity while the concept of Random Oversampling primarily involves randomly sampling and duplicating samples from the minority class into the dataset to achieve class balance by ensuring that all classes have the same quantity. Moreover, SMOTE (Synthetic Minority Oversampling Technique; Chawla et al., 2002) is a method derived as an improvement upon Random Oversampling. In contrast to Random Oversampling, which involves randomly sampling and duplicating samples from the minority class, SMOTE generates synthetic new samples by considering the distance between minority class samples and their k nearest neighbors, and then introduces these synthetic samples into the dataset. Moreover, SMOTEENN (Menardi & Torelli, 2014) is an approach to address data imbalance by combining Undersampling and Oversampling techniques. Specifically, Edited Nearest Neighbors (ENN) and SMOTE belong to Undersampling and Oversampling techniques, respectively. SMOTEENN initially employs ENN for Undersampling to reduce the sample quantity of the majority class, followed by using SMOTE for oversampling to augment the sample quantity of the minority class. This dual strategy helps address both overfitting and underfitting issues.

3.3. Machine learning models

According to Barboza's et al. (2017) and Chen's et al. (2023) empirical results, Random Forest and XGBoost algorithm have the best predictive power among various machine learning models. Hence, we follow Barboza's et al. (2017) and Chen et al. (2023) to introduce Random Forest and XGBoost algorithm as our main employed machine learning models. The details of these two algorithm are demonstrated as follows.

3.3.1. Random Forest

The Random Forest algorithm is a type of ensemble learning, specifically a Bagging model, composed of multiple decision trees. The ultimate prediction is determined through a voting process (Breiman, 2001). The implementation of the Random Forest model involves the following steps: (1) conducting repeated sampling to create subsets of the data; (2) independently modeling and predicting each subset; (3) aggregating the predictions from all trees using a multi-decision approach to determine the final classification result.

The decision-making process in the Random Forest algorithm involves several steps: (1) computing the information amount (Entropy) at each node for every level of the tree; (2) evaluating the information gain (IG) by subtracting the weighted average information of the nodes from the classified information; (3) choosing the feature with the higher information gain as the basis for categorization. Equation (1) illustrates the calculation of information gain (IG):

$$\text{Entropy(P)} = \ -\textstyle\sum_{i=1}^{x} p_i * \ \ Log^{p_i} \tag{1}$$

$$\text{IG(B)} = Entropy(P) - \ \sum_{j=1}^{k} \frac{|P_j|}{|P|} * Entropy(P_j)$$

In Eq. (1), Entropy(P) signifies the information content within the classified node P; x represents the number of categories in the set and p represents the proportion of each category in the set; IG(B) corresponds to the weighted average information amount before classification, subtracted from the information after classification. This is achieved by dividing the set P into k equal portions based on feature B.

The characteristics of Random Forest model are outlined as follows: (1) it is more efficient than individual decision trees and adept at handling missing values and outliers; (2) it exhibits reduced susceptibility to overfitting issues; (3) it is less constrained by the need for extensive hyperparameter tuning for making predictions.

3.3.2. XGBoost

The XGBoost (Extreme Gradient Boosting) algorithm, introduced by Chen and Guestrin in 2016, builds upon and enhances the Gradient Boosted Decision Tree (GBDT) model. Recognized for its high training efficiency and superior learning performance, XGBoost is prominently featured in machine learning literature. Operating on the principles of Boosting, XGBoost aggregates numerous weak classifiers to form a robust classifier. During the training phase, the algorithm boosts the data error weight of the existing classifier and continues to train new classifiers. This process enables the new classifier to learn the characteristics of misclassified data, leading to subsequent performance improvements through iterative training.

Based on the GBDT model, XGBoost addresses supervised learning problems by (1) combining numerous tree models to form a robust classifier and (2) utilizing gradient descent to minimize residuals during the decision-making process.

Additionally, to counteract potential overfitting, the XGBoost algorithm introduces a regularization term as a penalty.

Equation (2) presents our employed tree model, namely Classification and Regression Tree (CART) model. In Eq. (2), $\hat{y_i}$, N , and f_n indicate the prediction result of the model, the total number of decision trees, and the nth decision tree.

$$\hat{y}_i = \sum_{n=1}^{N} f_n(x_i) \tag{2}$$

Equation (3) demonstrates the objective function of XGBoost, composed of loss function ($l(y_i, \hat{y}_i{}^{(p-1)} + f_p(x_i))$) and regularization term ($\Omega(f_p)$), respectively. In Eq. (3), loss function measures the difference between the real result and the predicted result and its purpose is to correct the residuals of each tree in the past and the residuals of the newly added trees. In addition, regularization term is used to solve the overfitting problem and the problem can be effectively prevented by adjusting the penalty value and controlling the hyperparameters γ and λ.

$$Obj^{(p)} = \sum_{i=1}^{k} l(y_i, \hat{y}_i{}^{(p-1)} + f_p(x_i)) + \Omega(f_p) \tag{3}$$

$$\text{Where } \Omega\left(f_p\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \text{ ,}$$

Where *Obj* represents the objective function; $\hat{y}_i$ and $\hat{y}_i{}^{(p-1)}$ stands for the value of the current tree model ($f_p(x_i)$) and the predicted result of the previous tree, respectively; *x* , *k*, *p*, *T*, and *w* represent input variables, the number of samples in the training set, the number of trees constructed, the size of the tree (namely the number of leaf nodes), and the weight of the leaf nodes, respectively. γ and λ are hyperparameters.

3.4. Confusion Matrix

Following the previous studies in machine learning models (e.g. Chen et al., 2023), we employ the confusion matrix to evaluate the model effectiveness of the classification and prediction results. In the confusion matrix, the prediction results can be divided into four groups: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). In this study, we call Positive (Negative) as firm bankruptcy (non-bankruptcy).[1] We define Positive (Negative) as firm bankruptcy (non-bankruptcy). The Confusion Matrix is illustrated in Table 4.

[Insert Table 4 in here]

The evaluation variables related to confusion matrix include F1-Score, AUCs (Area Under the ROC Curve), Type I Error, and Type II Error. The F1-Score is a weighted average of the precision rate and recall rate, shown as Eq. (4). The precision rate indicates how many of the samples predicted to be positive (negative) by the model are actually positive (negative) samples, namely $\frac{TP}{TP+FP}$ $(\frac{TN}{TN+FN})$. The recall rate indicates how many positive (negative) samples the model can successfully predict from actually positive (negative) samples, namely $\frac{TP}{TP+FN}$ $(\frac{TN}{TN+FP})$. In addition, we define the Type I Error (Type II Error) as the ratio of bankrupt (non-bankrupt) firm samples are misjudged as non-bankrupt (bankrupt) firms to the actual bankrupt (non-bankrupt) firm sample, namely $\frac{FN}{TP+FN}$ $(\frac{FP}{TN+FP})$.

$$\mathrm{F1-Score} = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{4}$$

It is important to note that, due to the relatively low number of bankruptcy observations in the testing set (accounting for only 0.37% of the testing sample), the True Positive (TP) rate is susceptible to significant fluctuations with a small sample

[1] In this study, since Positive means bankruptcy and Negative means non-bankruptcy, True Positive means an actual bankrupt firm and the prediction result is also bankrupt firm; True Negative represents an actual non-bankrupt firm and the prediction result is also non-bankrupt firm; False Positive represents an actual non-bankrupt firm and the prediction result is a bankrupt firm; False Negative represents an actual bankrupt firm and the prediction result is a non-bankrupt firm.

size (due to the smaller denominator in TP rate). To avoid excessive variability in performance metrics caused by the small sample size, such as a substantial and less stable fluctuation in Area Under the Curve (AUC) where TP rate is on the vertical axis (i.e., Recall: companies actually bankrupt but correctly predicted as bankrupt), this study takes into consideration both F1-score and AUC as performance metrics for predicting model effectiveness.

We expect that including the social network characteristics variables of CEOs and other managers are able to improve the F1-Score, AUCs (Area Under the ROC Curve), Type I Error, and Type II Error of the machine learning models in addition to the Barboza's et al. (2017) financial variables.

4. Empirical Analyses

This study utilizes a sample of bankrupt and non-bankrupt U.S. firms from the year 2000 to 2020. Building upon the baseline model settings with 11 financial variables proposed by Barboza et al. (2017), the research investigates whether including 20 social network characteristic variables related to CEOs and other managers can significantly enhance the bankruptcy prediction performance. In addition, to ensure the robustness of the model results, this study conducts predictions for each machine learning model using 30 sets of random states and analyzes the model prediction effectiveness. The prediction performance is then presented by averaging the performance metrics across the 30 sets of random states for each model.

Tables 5 and 6 present the descriptive statistics for the social network characteristic variables and financial variables in the sample of this study, respectively. As the results presented here are not normalized, the occurrence of extreme values is observed. It is noteworthy that the machine learning models employed in this study are tree-based, and as such, the empirical results are not influenced by extreme values. Moreover, from

Table 5, it can be observed that the variables with larger numerical scales among the variables of CEOs' and other managers' social network characteristics are features belonging to Betweenness and Degree categories. Meanwhile, from Table 6, variables with larger numerical scales among the financial variables include EMV_Debt, ROE_Chg, PB_Chg, and Sales_growth, among others.

[Insert Table 5 in here]

[Insert Table 6 in here]

Table 7 presents the results of the model prediction effectiveness for corporate bankruptcy in the next one to six years before and after incorporating CEOs' and other managers' social network characteristic variables. From Panel A of Table 7, it is evident that in terms of predicting corporate bankruptcies in the next one year, the inclusion of CEOs' and other managers' social network characteristic variables indeed enhances the prediction effectiveness (e.g., F1-score) of both Random Forest (RF) and Extreme Gradient Boosting (XGBoost) machine learning models when combined with various data imbalance processing methods. Notably, the XGBoost model, when combined with the RandomOverSampler data imbalance processing method, shows the highest improvement in performance (e.g., its F1-score increases from 0.0819 to 0.3752). When considering AUC, the RF or XGBoost models, when combined with the EasyEnsemble data imbalance processing method, exhibit the highest performance (with AUC values of 0.8712 and 0.8688, respectively). This represents an increase compared to the AUC under the scenario where only financial variables are used as model input variables (with AUC values of 0.8274 and 0.8285, respectively).

Furthermore, this study also finds that when RF is combined with the EasyEnsemble data imbalance processing method, the number of False Negatives (FN) – instances where the model predicts non-bankruptcy, but the actual status is bankruptcy (i.e. Type I error) – slightly increases from 0.2333 to 0.7778. Simultaneously, the number of False Positives (FP) – instances where the model predicts bankruptcy, but the actual status is non-bankruptcy (i.e. Type II error) – significantly decreases from 504.3000 to 253.0611. Additionally, similar performance patterns are found when XGBoost is combined with the EasyEnsemble data imbalance processing method, with FN increasing from 0.3056 to 1.0778 and FP decreasing from 482.0945 to 149.9444. These results indicate that CEOs' and other managers' social network characteristics variables can substantially improve the number of False Positives (Type II error) while maintaining a certain level of False Negatives (Type I error). This implies that these variables can reduce the likelihood of misjudging non-bankrupt firms as bankrupt ones. It also suggests that information derived from CEOs' and other managers' social network characteristics variables can increase the chances of banks providing credit to truly healthy clients, increase the efficiency of fund utilization, and effectively improve credit management performance.

Panel B presents the results of corporate bankruptcy prediction effectiveness for the next two years. The empirical results indicate that, with the inclusion of CEOs' and other managers' social network characteristics variables, both RF and XGBoost show improvements in various data imbalance processing methods when evaluated based on the F1-score. Particularly noteworthy is the highest performance of XGBoost combined with RandomOverSampler, where its F1-score increases from 0.0724 to 0.3279. When evaluating based on AUC, the RF or XGBoost models combined with EasyEnsemble show only marginal increases compared to using only financial variables as model input variables. For instance, the former's AUC increases from 0.8206 to 0.8332, and the latter's AUC increases from 0.8192 to 0.8303. In comparison to the corporate bankruptcy prediction performance for the next one year, the benefits of CEOs' and other managers' social network characteristics variables are somewhat diminished for predicting corporate bankruptcy over the next two years.

Panels C to F present the results of corporate bankruptcy prediction effectiveness for the next three to six years. The empirical results indicate that, with the inclusion of CEOs' and other managers' social network characteristics variables, both RF and XGBoost show improvements in various data imbalance processing methods when evaluated based on the F1-score. Particularly noteworthy is the highest performance of XGBoost combined with RandomOverSampler. When evaluating based on AUC, the RF or XGBoost models combined with EasyEnsemble exhibit a slight decrease compared to using only financial variables as model input variables, indicating that the contribution of CEOs' and other managers' social network characteristics variables to long-term bankruptcy prediction performance is not as significant as in the short-term bankruptcy prediction performance.

[Insert Table 7 in here]

To reinforce the significance of the incremental enhancement in the effectiveness of bankruptcy prediction models after incorporating CEOs' and other managers' social network characteristics variables, this study conducts mean difference tests on F1-score and AUC for each machine learning model under 30 sets of random states before and after adding these SN_C variables, as shown in Tables 8 and 9. The results of Table 8 present that F1-score significantly increase after including CEOs' and other managers' social network characteristics variables, especially the best performance observed with XGBoost combined with RandomOverSampler, which exhibits the highest level of improvement. In addition, upon examining the bankruptcy prediction results for the next one to six years, the study finds that short-term prediction performance levels and improvements are generally more favorable than those for the long-term prediction performance levels and improvements. This suggests that CEOs' and other managers' social networks characteristic variables not only contribute to improving bankruptcy prediction effectiveness but also have different impacts on the prediction performance across different time horizons (namely bankruptcy term structure).

[Insert Table 8 in here]

Table 9 presents the comparative analyses results of AUCs before and after incorporating CEOs' and other managers' social network characteristics variables. The empirical results indicate that, in terms of bankruptcy prediction effectiveness for the next one year, the inclusion of these variables almost consistently leads to a significant increase in AUCs across various model settings (except for RF combined with the BalancedBagging model). Notably, when RF and XGBoost are combined with data imbalance processing methods such as RandomUnderSampling or EasyEnsemble, they exhibit relatively higher AUCs and a more substantial improvement. For instance, the AUC for RF (XGBoost) combined with RandomUnderSampling increases from 0.8124 (0.7875) to 0.8627 (0.8610), and the AUC for RF (XGBoost) combined with EasyEnsemble increases from 0.8274 (0.8285) to 0.8712 (0.8688), showing a significant level of improvement.

In terms of bankruptcy prediction effectiveness for the next two years, after incorporating CEOs' and other managers' social network characteristics variables, the AUCs for RF (XGBoost) combined with EasyEnsemble increases from 0.8206 (0.8192) to 0.8332 (0.8303), with an increment of 0.0126 (0.0111). While the increase in AUC is statistically significant, the magnitude of improvement is relatively smaller compared to the increase observed for the next one year. Furthermore, for the bankruptcy prediction effectiveness over the next three to six years, the inclusion of CEOs' and other managers' social network characteristic variables does not lead to a significant increase in AUCs for RF (XGBoost) combined with EasyEnsemble. This indicates that

CEO and other managers' social network characteristics variables primarily contribute to enhancing the short-term effectiveness of bankruptcy prediction. However, as the prediction horizon extends, the model's ability to effectively improve bankruptcy prediction performance gradually diminishes.

[Insert Table 9 in here]

To provide a reasonable explanation for the above prediction results, this study employed the Random Forest method to conduct feature selection for Barboza et al.'s (2017) 11 financial variables and the proposed 20 CEOs' and other managers' social network characteristics variables over different training periods. The empirical results reveal that the "Betweenness_CEO" variable (indicating the degree to which the CEO controls the flow of information) exhibits the highest importance among the social network characteristics variables, ranking fourth in importance among all 31 variables. Following closely is the "std_Degree_nCEO" variable (indicating the standard deviation of the social network size among other managers), which ranks approximately sixth in importance among all 31 variables. Since the "CEO Betweenness" variable represents the CEO's ability to control the speed of information flow (i.e., communication value), it affects the level of incomplete information within the firm. According to Duffie and Lando's (2001) structural-form credit risk model, incomplete information has a stronger explanatory power for short-term credit risk. Therefore, the finding of this study that CEOs and other managers' social network characteristics variables significantly improve the effectiveness of short-term bankruptcy prediction aligns with the views of Duffie and Lando (2001) and Yu (2005), who emphasize the substantial impact of incomplete information on short-term credit risk assessment.

[Insert Table 10 in here]

5. Conclusions

This study primarily relies on machine learning models to empirically investigate whether including CEOs' and other managers' social network characteristics variables provides additional bankruptcy prediction effectiveness and explanatory aspects beyond the existing financial variables proposed by Barboza et al. (2017). In recent twenty years, bankruptcy prediction has gained significant attentions from corporations, external investors, and financial institutions. Concurrently, the application of machine learning techniques in financial and accounting research has flourished. Both academic and practical communities recognize the effectiveness of machine learning models in bankruptcy prediction. However, there is still a need for further strengthening the explanatory aspects. Therefore, this study, while focusing on enhancing bankruptcy prediction effectiveness, places a greater emphasis on exploring the explanatory dimensions, theoretical foundations, and economic implications of CEOs' and other managers' social network characteristics variables in bankruptcy prediction.

The empirical results of this study demonstrate that, using the financial variables proposed by Barboza et al. (2017) as the baseline model, the inclusion of CEOs' and other managers' social network characteristics variables significantly enhances the F1-score of the model, both in the short term and long term. The improvement is particularly notable in the short-term predictions. However, concerning the AUC as a model prediction effectiveness indicator, the impacts of CEOs' and other managers' social network characteristics variables are significant only for short-term bankruptcy prediction, especially for the prediction of bankruptcy in the next one year. To provide a reasonable explanation for this phenomenon, the study conducts feature selection and finds that the degree of CEO control over information flow (i.e. Betweenness_CEO) has the highest information content among the social network characteristics variables regarding corporate bankruptcy. Since the CEO's control over information flow represents the manipulation of the speed of information transmission (i.e., communication value), it influences the level of incomplete information within the firm. Hence, the significant improvement in short-term bankruptcy prediction effectiveness observed with CEOs' and other managers' social network characteristics variables aligns with the perspectives of Duffie and Lando (2001) and Yu (2005), who emphasize the substantial impact of incomplete information on short-term credit risk assessment.

Furthermore, this study also finds that CEOs' and other managers' social network characteristics variables can significantly improve the number of False Positives (Type II error) while maintaining a certain number of False Negatives (Type I error), particularly in short-term bankruptcy prediction. This implies that CEOs' and other managers' social network characteristics variables can reduce the likelihood of misjudging non-bankrupt firms as bankrupt ones. This also suggests that the information provided by CEOs' and other managers' social network characteristics variables increases the opportunity for banks to extend credit to financially sound clients, enhances the efficiency of fund utilization, and effectively improves credit management performance. Hence, the CEOs' and other managers' social network characteristics variables can help prevent or mitigate economic losses resulting from corporate bankruptcy. Finally, for financial institutions and regulatory authorities, leveraging information about the social network structure characteristics of top management team members can aid in assessing whether a firm has bankruptcy concerns, facilitating informed lending decisions or regulatory measures.

References

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, 94, 164–184.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609.

Anderson, A., Park, J., & Jack, S. (2007). Entrepreneurial social capital: Conceptualizing social capital in new high-tech firms. *International Small Business Journal*, 25(3), 245–272.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71–111.

Bloch, F., Genicot, G., & Ray, D. (2008). Informal insurance in social networks. *Journal of Economic Theory*, 143(1), 36–58.

Bonsall IV, S. B., & Miller, B. P. (2017). The impact of narrative disclosure readability on bond ratings and the cost of debt. *Review of Accounting Studies* 22 (2), 608-643.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, 27(1), 55–71.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Burt, R. S. (2001). Structural holes versus network closure as social capital. *Social Capital: Theory and Research*, 31-56.

Carton, R., & Hofer, C. (2006). Measuring organizational performance. Edward Elgar Publishing.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, T.K., & Tseng, Y. (2021). Readability of notes to consolidated financial statements and corporate bond yield Spread. *European Accounting Review* 30(1), 83-113

Chen, T.K., Tseng, Y. (2022). Management team categories, social network characteristics, and corporate credit risk. *Working paper*, National Yang Ming Chiao Tung University.

Chen, T.K., Liao, H.H., Chen, G.D., Kang, W.H., & Lin, Y.C. (2023a). Bankruptcy

Prediction Using Machine Learning Models with the Text-based Communicative Value of Annual Reports. *Expert Systems with Applications,* 233*,* 120714.

Chen, T.K., Chang, T.R., & Lin, Y.C. (2023b). The uncertainty of text-based communicative value of annual reports and bankruptcy prediction. *Working paper*, National Yang Ming Chiao Tung University.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94*, S95-S120.

Duffie, D., & Lando, D. (2001). Term structures of credit spreads with incomplete accounting information. *Econometrica*, 69(3), 633-664.

El-Khatib, R., Fogel, K., & Jandik, T. (2015). CEO network centrality and merger performance. *Journal of Financial Economics*, 116(2), 349–382.

Ertugrul, M., Lei, J., Qiu, J., & Wan, C. (2017). Annual report readability, tone ambiguity, and the cost of borrowing. *Journal of Financial and Quantitative Analysis*, 52(2), 811–836.

Fan, Y., Boateng, A., Ly, K. C., & Jiang, Y. (2021). Are bonds blind? Board-CEO social networks and firm risk. *Journal of Corporate Finance*, 68, 101922.

Ferris, S. P., Javakhadze, D., & Rajkovic, T. (2017). CEO social capital, risk-taking and corporate policies. *Journal of Corporate Finance*, 47, 46–71.

Ferris, S. P., Javakhadze, D., & Rajkovic, T. (2019). An international analysis of CEO social capital and corporate risk-taking. *European Financial Management*, 25(1), 3–37.

Hido, S., Kashima, H., & Takahashi, Y. (2009). Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 2*(5-6), 412-426.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429-449.

Liang, D., Lu, C. C., Tsai, C. F., & Shih, G. A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2), 561-572.

Lin, C.Y., Chen, Y., Ho, P.H., Yen, J.F. (2020). CEO overconfidence and bank loan contracting. *Journal of Corporate Finance* 64, 101637.

Kim, A., & Yoon, S. (2021). Corporate bankruptcy prediction with domain-adapted BERT. 2021 EMNLP, 3rd Workshop on ECONLP.

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research* 274, 743–758.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2), 449-470.

Malmendier, U., & Tate, G. (2005). CEO overconfidence and corporate investment. *The Journal of Finance*, *60*(6), 2661–2700

Malmendier, U., Tate, G., & Yan, J. (2011). Overconfidence and early-life experiences: the effect of managerial traits on corporate financial policies. *The Journal of Finance, 66*(5), 1687-1733.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery, 28*, 92-122.

Nahapiet, J., & Ghoshal, S. (1998). Social Capital, Intellectual Capital, and the Organizational Advantage. *Academy of Management Review*, 23(2), 242–266.

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

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*(2), 464-473.

Seebeck, A. & Kaya, D. (2023). The power of words: An empirical analysis of the communicative value of extended auditor reports. *European Accounting Review* 32 (5), 1185-1215.

Sun, J., Li, H., Huang, Q. H., & He, K. Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, 57, 41–56.

Xing, J., Zhang, Y., & Xiong, X. (2023). Social capital, independent director connectedness, and stock price crash risk. *International Review of Economics & Finance*, 83, 786–804.

Zhang, L., Peng, F., Shan, Y. G., & Chen, Y. (2023). CEO social capital and litigation risk. *Finance Research Letters*, 51, 103405.

Table 1. Variables Definitions: CEOs' and Other Managers' Social Network Characteristics Information

Variable	Definitions
Degree_CEO	In a network, the number of direct contacts a CEO has with other managers.
Betweenness_CEO	The frequency with which the CEO is on the shortest path between two other managers in the network
Closeness_CEO	The reciprocal of the sum of the shortest distances between the CEO and all other managers.
Eigenvector_CEO	CEO's importance in the network
Degree_nCEO	The average number of direct contacts between non-CEO managers and other managers in the network.
Betweenness_nCEO	In the network, the shortest path between a non-CEO manager and two other managers
Closeness_nCEO	The reciprocal of the sum of the shortest distances between non-CEO managers and all other managers
Eigenvector_nCEO	Average importance of non-CEO managers in the network
Std_Degree_nCEO	The standard deviation of the Degree of all non-CEO managers in the same company
Std_Betweenness_nCEO	The standard deviation of the Betweenness of all non-CEO managers in the same company
Std_Closeness_nCEO	The standard deviation of the Closeness of all non-CEO managers in the same company
Std_Eigenvector_nCEO	The standard deviation of the Eigenvector of all non-CEO managers in the same company
Degree_All	The average number of direct ties that all managers have with other managers in the network
Betweenness_All	In a network, the average number of all managers on the shortest path between two other managers
Closeness_All	The average number of reciprocals of the sum of the shortest distances between all managers and all other managers
Eigenvector_All	The average importance of all managers in the network
Std_Degree_All	The standard deviation of the Degree of all managers in the same company
Std_Betweenness_All	The standard deviation of the Betweenness of all managers in the same company
Std_Closeness_All	The standard deviation of the Closeness of all managers in the same company
Std_Eigenvector_All	The standard deviation of the Eigenvector of all managers in the same company

Table 2. Variables Definitions: Barboza's et al. (2017) Financial Variables

Variable	Definitions
NWC_TA	Net working capital divided by total assets
RE_TA	Retained earnings divided by total assets
EBIT_TA	Earnings before interests and taxes divided by assets
EMV_Debt	Equity market value divided by total debts
Sales_TA	Net sales divided by total assets
EBIT_Sales	Earnings before interests and taxes divided by net sales
TA_growth	The yearly growth rate of total assets
Sales_growth	The yearly growth rate of net sales
EMP_growth	The yearly growth rate of the number of employee
ROE_Chg	The annual change of return on equity
PB_Chg	The annual change of equity market-to-book value ratio

Note: The first five variables are Altman's (1968) Z score variables; the last six variables are from Carton and Hofer (2006).

Table 3. Sample Distributions

Year	Non-Default	Default
2000	725	9
2001	929	13
2002	986	13
2003	1410	22
2004	1538	30
2005	1619	37
2006	1604	35
2007	1628	35
2008	1608	22
2009	1586	19
2010	1552	22
2011	1535	21
2012	1562	19
2013	1544	18
2014	1547	13
2015	1609	10
2016	1590	9
2017	1583	7
2018	1575	4
2019	1746	4
2020	1685	2

Table 4. Confusion Matrix

		Predicted	
		Bankrupt	Non-Bankrupt
Actual	Bankrupt	True Positive (TP)	False Negative (FN)
	Non-Bankrupt	False Positive (FP)	True Negative (TN)

Table 5. Summary Statistics of Major Variables: CEOs' and Other Managers' Social Network Characteristics

Variable	Obs	Mean	Std	Min	Max
Degree_CEO	31525	335.146	648.297	3.000	7255.000
Betweenness_CEO	31525	940888.800	2367984.708	0.000	47955381.682
Closeness_CEO	31525	0.000	0.006	0.000	0.333
Eigenvector_CEO	31525	0.007	0.054	0.000	0.999
Degree_nCEO	31525	394.816	419.530	3.000	5057.692
Betweenness_nCEO	31525	1745252.466	1868974.109	0.000	19028718.023
Closeness_nCEO	31525	0.000	0.006	0.000	0.333
Eigenvector_nCEO	31525	0.007	0.025	0.000	0.792
std_Degree_nCEO	31525	437.319	447.393	0.000	3117.102
std_Betweenness_nCEO	31525	2727402.459	3075692.442	0.000	28494824.899
std_Closeness_nCEO	31525	0.000	0.000	0.000	0.018
std_Eigenvector_nCEO	31525	0.015	0.046	0.000	0.470
Degree_all	31525	387.571	414.741	3.000	5060.077
Betweenness_all	31525	1657101.196	1754705.700	0.000	16676334.088
Closeness_all	31525	0.000	0.006	0.000	0.333
Eigenvector_all	31525	0.007	0.025	0.000	0.811
std_Degree_all	31525	433.921	435.137	0.000	2833.120
std_Betweenness_all	31525	2668795.165	2945083.055	0.000	26416517.896
std_Closeness_all	31525	0.000	0.000	0.000	0.019
std_Eigenvector_all	31525	0.015	0.045	0.000	0.454

Note: The above 20 variables are the descriptive statistics of the characteristic variables of managers' social network information used in this study. After data pre-processing, no missing values are included, with a total of 31,525 sample observations. See Table 1 for the definitions of various variables.

Table 6. Summary Statistics of Major Variables: Barboza's et al. (2017) Financial Variables

Variable	Obs	Mean	Std	Min	Max
NWC_TA	31525	0.197	0.217	-8.585	0.954
RE_TA	31525	-0.149	1.902	-111.250	2.529
EBIT_TA	31525	0.055	0.202	-14.609	1.745
EMV_Debt	31525	433.654	8637.975	0.000	750555.500
Sales_TA	31525	0.997	0.775	-0.025	15.961
EBIT_Sales	31525	-0.862	24.126	-2162.100	119.359
TA_growth	31525	0.141	0.507	-0.894	30.314
Sales_growth	31525	0.460	23.482	-9.286	3701.467
EMP_growth	31525	0.117	3.762	-0.993	629.500
ROE_Chg	31525	0.753	130.781	-786.654	23152.310
PB_Chg	31525	-0.637	91.791	-6996.573	5600.936

Note: The above 11 variables are the descriptive statistics of financial variables used by Barboza et al. (2017). After data pre-processing, no missing values are included, with a total of 31,525 sample observations. The definitions of various variables are detailed in Table 2.

Table 7. Comparative Analyses of the Effectiveness of Bankruptcy Prediction Models (Forecasting period is the next one to six years)

Panel A. Forecasting period is the next one year

A.1. FIN variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	EasyEnsemble	0.8274	0.0232	5.7667	1127.0330	504.3000	0.2333
	BalancedBagging	0.8454	0.0316	5.4778	1281.5280	349.8055	0.5222
	RandomUnderSampling	0.8124	0.0227	5.6056	1129.4670	501.8667	0.3944
	RandomOverSampler	0.5172	0.0619	0.3278	1631.1670	0.1667	5.6722
	SMOTE	0.6227	0.0864	2.1167	1591.6280	39.7056	3.8833
	SOMTEENN	0.6631	0.0708	2.8167	1563.0610	68.2722	3.1833
XGBoost	EasyEnsemble	0.8285	0.0240	5.6944	1149.2390	482.0945	0.3056
	BalancedBagging	0.8163	0.0310	5.1889	1292.9830	338.3500	0.8111
	RandomUnderSampling	0.7875	0.0219	5.3389	1137.3170	494.0167	0.6611
	RandomOverSampler	0.5454	0.0819	0.8111	1622.0280	9.3056	5.1889
	SMOTE	0.6458	0.0507	2.6000	1538.7500	92.5833	3.4000
	SOMTEENN	0.6785	0.0514	3.3056	1512.7830	118.5500	2.6944

A.2. FIN&SN_C variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	**EasyEnsemble**	**0.8712**	0.0390	5.2222	1378.2720	253.0611	0.7778
	BalancedBagging	0.8282	0.0470	4.5389	1454.1500	177.1833	1.4611
	RandomUnderSampling	0.8627	0.0386	5.1389	1379.1720	252.1611	0.8611
	RandomOverSampler	0.5434	0.1454	0.7389	1631.3170	0.0167	5.2611
	SMOTE	0.6715	**0.2513**	2.1000	1624.2720	7.0611	3.9000
	SOMTEENN	0.6897	0.2283	2.3056	1620.7060	10.6278	3.6944
XGBoost	**EasyEnsemble**	**0.8688**	0.0594	4.9222	1481.3890	149.9444	1.0778
	BalancedBagging	0.8358	0.0702	4.5778	1515.6500	115.6833	1.4222
	RandomUnderSampling	0.8610	0.0534	4.9500	1459.8440	171.4889	1.0500
	RandomOverSampler	0.6461	**0.3752**	2.1167	1630.5560	0.7778	3.8833
	SMOTE	0.7071	0.3397	2.8167	1625.7780	5.5556	3.1833
	SOMTEENN	0.7112	0.2997	2.8611	1623.6390	7.6944	3.1389

Panel B. Forecasting period is the next two years

B.1. FIN variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	EasyEnsemble	0.8206	0.0232	11.5400	2247.5670	1008.8330	0.4600
	BalancedBagging	0.8427	0.0317	11.0267	2559.7270	696.6733	0.9733
	RandomUnderSampling	0.8090	0.0227	11.2733	2251.2330	1005.1670	0.7267
	RandomOverSampler	0.5132	0.0490	0.4667	3255.9800	0.4200	11.5333
	SMOTE	0.6172	0.0769	3.7533	3174.8270	81.5733	8.2467
	SOMTEENN	0.6691	0.0684	5.3133	3116.6130	139.7867	6.6867
XGBoost	EasyEnsemble	0.8192	0.0239	11.3400	2293.3870	963.0133	0.6600
	BalancedBagging	0.8149	0.0310	10.3733	2581.9130	674.4867	1.6267
	RandomUnderSampling	0.7821	0.0216	10.5933	2263.8070	992.5933	1.4067
	RandomOverSampler	0.5397	0.0724	1.3333	3238.1070	18.2933	10.6667
	SMOTE	0.6472	0.0475	4.8000	3067.0530	189.3467	7.2000
	SOMTEENN	0.6838	0.0484	6.1667	3013.5530	242.8467	5.8333

B.2. FIN&SN_C variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	**EasyEnsemble**	**0.8332**	0.0365	9.6667	2757.5730	498.8267	2.3333
	BalancedBagging	0.7810	0.0436	8.2733	2908.6200	347.7800	3.7267
	RandomUnderSampling	0.8250	0.0364	9.5867	2758.6470	497.7533	2.4133
	RandomOverSampler	0.5366	0.1289	1.1400	3256.3330	0.0667	10.8600
	SMOTE	0.6419	**0.2109**	3.5333	3240.6730	15.7267	8.4667
	SOMTEENN	0.6638	0.1926	4.0400	3232.5400	23.8600	7.9600
XGBoost	**EasyEnsemble**	**0.8303**	0.0548	9.1600	2952.9670	303.4333	2.8400
	BalancedBagging	0.8063	0.0652	8.5467	3023.4600	232.9400	3.4533
	RandomUnderSampling	0.8299	0.0502	9.3667	2909.6530	346.7467	2.6333
	RandomOverSampler	0.6208	**0.3279**	3.4533	3254.3930	2.0067	8.5467

	SMOTE	0.6910	0.3079	4.8267	3243.4530	12.9467	7.1733
	SOMTEENN	0.7005	0.2740	4.9667	3238.5330	17.8667	7.0333

Panel C. Forecasting period is the next three years

C.1. FIN variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	EasyEnsemble	0.8228	0.0226	17.1083	3349.2170	1510.7830	0.6417
	BalancedBagging	0.8493	0.0310	16.3667	3818.5670	1041.4330	1.3833
	RandomUnderSampling	0.8094	0.0221	16.6000	3357.8000	1502.2000	1.1500
	RandomOverSampler	0.5126	0.0475	0.5833	4859.6580	0.3417	17.1667
	SMOTE	0.6062	0.0672	4.9417	4736.5580	123.4417	12.8083
	SOMTEENN	0.6723	0.0647	7.7333	4646.4330	213.5667	10.0167
XGBoost	EasyEnsemble	0.8144	0.0230	16.5583	3421.6000	1438.4000	1.1917
	BalancedBagging	0.8172	0.0300	15.1833	3854.9330	1005.0670	2.5667
	RandomUnderSampling	0.7745	0.0208	15.3917	3375.6500	1484.3500	2.3583
	RandomOverSampler	0.5382	0.0684	1.7417	4833.3830	26.6167	16.0083
	SMOTE	0.6444	0.0435	6.7250	4571.0170	288.9833	11.0250
	SOMTEENN	0.6890	0.0455	8.9167	4489.7500	370.2500	8.8333

C.2. FIN&SN_C variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	EasyEnsemble	0.7907	0.0331	12.9750	4117.9500	742.0500	4.7750
	BalancedBagging	0.7531	0.0402	11.2917	4341.2580	518.7416	6.4583
	RandomUnderSampling	**0.7946**	0.0337	13.1583	4115.9000	744.1000	4.5917
	RandomOverSampler	0.5390	0.1393	1.6583	4859.8670	0.1333	16.0917
	SMOTE	0.6176	**0.1766**	4.5083	4833.9420	26.0583	13.2417
	SOMTEENN	0.6446	0.1687	5.4583	4821.1500	38.8500	12.2917
XGBoost	EasyEnsemble	0.8014	0.0499	12.6417	4395.1080	464.8917	5.1083
	BalancedBagging	0.7912	0.0599	11.9250	4500.1670	359.8333	5.8250
	RandomUnderSampling	**0.8091**	0.0465	13.1667	4327.7170	532.2833	4.5833
	RandomOverSampler	0.6184	**0.3228**	4.7083	4856.4830	3.5167	13.0417
	SMOTE	0.6721	0.2710	6.5167	4839.0500	20.9500	11.2333
	SOMTEENN	0.6820	0.2471	6.7750	4831.7750	28.2250	10.9750

Panel D. Forecasting period is the next four years

D.1. FIN variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	EasyEnsemble	0.8235	0.0221	22.8000	4441.5110	2038.4890	0.8667
	BalancedBagging	0.8477	0.0301	21.7000	5071.8550	1408.1440	1.9667
	RandomUnderSampling	0.8115	0.0217	22.1444	4465.2000	2014.8000	1.5222
	RandomOverSampler	0.5143	0.0542	0.7778	6479.8330	0.1667	22.8889
	SMOTE	0.6211	0.0661	6.6667	6311.7330	168.2667	17.0000
	SOMTEENN	0.6910	0.0636	10.5556	6185.3670	294.6333	13.1111
XGBoost	EasyEnsemble	0.8092	0.0222	21.7778	4542.9890	1937.0110	1.8889
	BalancedBagging	0.8118	0.0287	19.9333	5123.3780	1356.6220	3.7333
	RandomUnderSampling	0.7724	0.0202	20.3556	4492.6450	1987.3560	3.3111
	RandomOverSampler	0.5401	0.0690	2.2778	6444.1110	35.8889	21.3889
	SMOTE	0.6472	0.0419	8.8222	6090.3780	389.6222	14.8444
	SOMTEENN	0.7025	0.0443	11.9000	5977.4670	502.5333	11.7667

D.2. FIN&SN_C variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	EasyEnsemble	0.7735	0.0318	16.5889	5483.7890	996.2111	7.0778
	BalancedBagging	0.7487	0.0393	14.6556	5782.6670	697.3333	9.0111
	RandomUnderSampling	**0.7781**	0.0325	16.8556	5484.9000	995.1000	6.8111
	RandomOverSampler	0.5448	**0.1621**	2.2111	6479.8220	0.1778	21.4556
	SMOTE	0.6101	0.1619	5.4889	6443.0890	36.9111	18.1778
	SOMTEENN	0.6368	0.1562	6.7667	6425.4440	54.5556	16.9000
XGBoost	EasyEnsemble	0.8013	0.0482	16.6444	5836.1000	643.9000	7.0222
	BalancedBagging	0.7860	0.0572	15.5000	5982.0560	497.9445	8.1667
	RandomUnderSampling	**0.8040**	0.0448	17.2111	5747.5890	732.4111	6.4556
	RandomOverSampler	0.6233	**0.3321**	6.0222	6474.6560	5.3444	17.6444
	SMOTE	0.6679	0.2577	8.1556	6449.4890	30.5111	15.5111
	SOMTEENN	0.6768	0.2327	8.5556	6439.2450	40.7556	15.1111

Panel E. Forecasting period is the next five years

E.1. FIN variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	EasyEnsemble	0.8187	0.0216	28.6833	5537.4170	2603.5830	1.3167
	BalancedBagging	0.8463	0.0293	27.4333	6324.8500	1816.1500	2.5667
	RandomUnderSampling	0.8082	0.0212	27.9833	5553.5000	2587.5000	2.0167
	RandomOverSampler	0.5193	0.0733	1.1667	8140.6000	0.4000	28.8333
	SMOTE	0.6264	0.0657	8.4833	7923.8670	217.1333	21.5167
	SOMTEENN	0.6956	0.0618	13.4000	7756.5000	384.5000	16.6000
XGBoost	EasyEnsemble	0.8048	0.0217	27.4000	5670.7830	2470.2170	2.6000
	BalancedBagging	0.8173	0.0283	25.4833	6396.2670	1744.7330	4.5167
	RandomUnderSampling	0.7767	0.0201	26.0167	5596.3500	2544.6500	3.9833
	RandomOverSampler	0.5481	0.0774	3.1667	8094.4830	46.5167	26.8333
	SMOTE	0.6581	0.0434	11.5167	7655.3000	485.7000	18.4833
	SOMTEENN	0.7113	0.0444	15.2167	7504.0670	636.9333	14.7833

E.2. FIN&SN_C variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	EasyEnsemble	0.7677	0.0314	20.8167	6872.6830	1268.3170	9.1833
	BalancedBagging	0.7649	0.0407	19.2333	7248.4000	892.6000	10.7667
	RandomUnderSampling	**0.7768**	0.0324	21.3000	6879.7330	1261.2670	8.7000
	RandomOverSampler	0.5445	**0.1612**	2.7167	8140.6000	0.4000	27.2833
	SMOTE	0.6085	0.1569	6.7000	8092.7000	48.3000	23.3000
	SOMTEENN	0.6372	0.1535	8.4833	8069.3330	71.6667	21.5167
XGBoost	EasyEnsemble	0.7914	0.0463	20.6167	7304.4830	836.5167	9.3833
	BalancedBagging	0.7835	0.0556	19.5000	7491.2830	649.7167	10.5000
	RandomUnderSampling	**0.8043**	0.0437	21.8833	7176.5670	964.4333	8.1167
	RandomOverSampler	0.6160	**0.3092**	7.1500	8133.0330	7.9667	22.8500
	SMOTE	0.6546	0.2357	9.5500	8099.6830	41.3167	20.4500
	SOMTEENN	0.6620	0.2077	10.0167	8084.8830	56.1167	19.9833

Panel F. Forecasting period is the next six years

F.1. FIN variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	EasyEnsemble	0.8121	0.0209	34.2667	5537.4170	3207.4000	1.7333
	BalancedBagging	0.8376	0.0281	32.6000	6324.8500	2255.0330	3.4000
	RandomUnderSampling	0.7996	0.0205	33.2333	5553.5000	3171.7670	2.7667
	RandomOverSampler	0.5189	0.0712	1.3667	8140.6000	0.8667	34.6333
	SMOTE	0.6299	0.0658	10.3333	7923.8670	267.5000	25.6667
	SOMTEENN	0.7120	0.0648	17.0000	7756.5000	471.9333	19.0000
XGBoost	EasyEnsemble	0.8000	0.0211	32.8000	5670.7830	3044.7670	3.2000
	BalancedBagging	0.8082	0.0270	30.1667	6396.2670	2169.0330	5.8333
	RandomUnderSampling	0.7764	0.0197	31.3667	5596.3500	3117.3000	4.6333
	RandomOverSampler	0.5581	0.0882	4.4000	8094.4830	59.8667	31.6000
	SMOTE	0.6764	0.0472	14.8333	7655.3000	579.2333	21.1667
	SOMTEENN	0.7194	0.0448	18.6667	7504.0670	779.5667	17.3333

F.2. FIN&SN_C variables

Model	Data Imbalanced Methods	AUC	F1-score	TP	TN	FP	FN
RF	EasyEnsemble	0.7824	0.0318	26.1667	6872.6830	1585.4330	9.8333
	BalancedBagging	0.7731	0.0402	23.8000	7248.4000	1124.4670	12.2000
	RandomUnderSampling	0.7868	0.0325	26.4000	6879.7330	1563.6000	9.6000
	RandomOverSampler	0.5472	0.1691	3.4000	8140.6000	0.7667	32.6000
	SMOTE	0.6044	0.1501	7.7333	8092.7000	59.4333	28.2667
	SOMTEENN	0.6311	0.1453	9.7667	8069.3330	88.6667	26.2333
XGBoost	EasyEnsemble	0.8001	0.0447	25.6000	7304.4830	1085.1000	10.4000
	BalancedBagging	0.7900	0.0531	24.0000	7491.2830	847.9000	12.0000
	RandomUnderSampling	**0.8112**	0.0416	27.0000	7176.5670	1248.9670	9.0000
	RandomOverSampler	0.6285	**0.3206**	9.3000	8133.0330	12.5667	26.7000
	SMOTE	0.6595	0.2218	11.7000	8099.6830	58.1000	24.3000
	SOMTEENN	0.6643	0.1980	12.1000	8084.8830	74.2667	23.9000

Note: AUC and F1-score are model performance measures. For the definitions of TP, TN, FP, and FN, please refer to the confusion matrix. Since the actual number of bankruptcies in the testing set of this study is small, TP rate is susceptible to large changes due to a small number of samples (because the denominator of TP rate is small), so the change range of AUC is also high (TP rate is its vertical axis). Due to the above considerations, this study also considers F1-score and AUC as the main performance indicators. In addition, FIN represents financial variables, and FIN&SN_C represents financial and social network characteristics variables.

Table 8. The Difference Tests Analyses in the Effectiveness of Bankruptcy Prediction Models: Before and After Introducing the Social Network Characteristics Variables (F1- score)

Panel A. Forecasting period is the next one year

Model	Data Imbalanced Methods	FIN	FIN&SN_C	Difference	T_stat	p-value
RF	EasyEnsemble	0.0232	0.0390	0.0158	39.9186	0.0000
	BalancedBagging	0.0316	0.0470	0.0155	21.9344	0.0000
	RandomUnderSampling	0.0227	0.0386	0.0159	35.4012	0.0000
	RandomOverSampler	0.0619	0.1454	0.0834	8.8234	0.0000
	SMOTE	0.0864	0.2513	0.1649	23.7204	0.0000
	SOMTEENN	0.0708	0.2283	0.1575	29.2072	0.0000
XGBoost	EasyEnsemble	0.0240	0.0594	0.0355	35.3443	0.0000
	BalancedBagging	0.0310	0.0702	0.0392	30.3682	0.0000
	RandomUnderSampling	0.0219	0.0534	0.0315	29.5957	0.0000
	RandomOverSampler	**0.0819**	**0.3752**	**0.2933**	**16.5613**	**0.0000**
	SMOTE	0.0507	0.3397	0.2889	23.1191	0.0000
	SOMTEENN	0.0514	0.2997	0.2483	24.5993	0.0000

Panel B. Forecasting period is the next two years

Model	Data Imbalanced Methods	FIN	FIN&SN	Difference	T_stat	p-value
RF	EasyEnsemble	0.0232	0.0365	0.0133	53.3574	0.0000
	BalancedBagging	0.0317	0.0436	0.0119	18.3899	0.0000
	RandomUnderSampling	0.0227	0.0364	0.0137	33.5666	0.0000
	RandomOverSampler	0.0490	0.1289	0.0800	11.5704	0.0000
	SMOTE	0.0769	0.2109	0.1341	28.7733	0.0000
	SOMTEENN	0.0684	0.1926	0.1242	34.3191	0.0000
XGBoost	EasyEnsemble	0.0239	0.0548	0.0310	40.3811	0.0000
	BalancedBagging	0.0310	0.0652	0.0342	35.4442	0.0000
	RandomUnderSampling	0.0216	0.0502	0.0286	29.2892	0.0000
	RandomOverSampler	**0.0724**	**0.3279**	**0.2555**	**19.4432**	**0.0000**
	SMOTE	0.0475	0.3079	0.2604	41.4762	0.0000
	SOMTEENN	0.0484	0.2740	0.2256	44.8410	0.0000

Panel C. Forecasting period is the next three years

Model	Data Imbalanced Methods	FIN	FIN&SN	Difference	T_stat	p-value
RF	EasyEnsemble	0.0226	0.0331	0.0105	47.2873	0.0000
	BalancedBagging	0.0310	0.0402	0.0092	20.0789	0.0000
	RandomUnderSampling	0.0221	0.0337	0.0116	32.7108	0.0000
	RandomOverSampler	0.0475	0.1393	0.0918	14.8795	0.0000
	SMOTE	0.0672	0.1766	0.1094	31.6430	0.0000
	SOMTEENN	0.0647	0.1687	0.1039	42.2221	0.0000
XGBoost	EasyEnsemble	0.0230	0.0499	0.0269	40.8625	0.0000
	BalancedBagging	0.0300	0.0599	0.0299	38.8438	0.0000
	RandomUnderSampling	0.0208	0.0465	0.0257	28.3858	0.0000
	RandomOverSampler	**0.0684**	**0.3228**	**0.2544**	**24.0816**	**0.0000**
	SMOTE	0.0435	0.2710	0.2275	40.5886	0.0000
	SOMTEENN	0.0455	0.2471	0.2015	42.8623	0.0000

Panel D. Forecasting period is the next four years

Model	Data Imbalanced Methods	FIN	FIN&SN	Difference	T_stat	p-value
RF	EasyEnsemble	0.0221	0.0318	0.0097	55.5339	0.0000
	BalancedBagging	0.0301	0.0393	0.0093	23.9203	0.0000
	RandomUnderSampling	0.0217	0.0325	0.0108	29.3556	0.0000
	RandomOverSampler	0.0542	0.1621	0.1079	23.6249	0.0000
	SMOTE	0.0661	0.1619	0.0959	27.0514	0.0000
	SOMTEENN	0.0636	0.1562	0.0927	38.4867	0.0000
XGBoost	EasyEnsemble	0.0222	0.0482	0.0261	59.7369	0.0000
	BalancedBagging	0.0287	0.0572	0.0285	48.6653	0.0000
	RandomUnderSampling	0.0202	0.0448	0.0246	29.7329	0.0000
	RandomOverSampler	**0.0690**	**0.3321**	**0.2631**	**38.3351**	**0.0000**
	SMOTE	0.0419	0.2577	0.2158	58.6306	0.0000

	SOMTEENN	0.0443	0.2327	0.1884	51.1958	0.0000

Panel E. Forecasting period is the next five years

Model	Data Imbalanced Methods	FIN	FIN&SN	Difference	T_stat	p-value
RF	EasyEnsemble	0.0216	0.0314	0.0099	45.5513	0.0000
	BalancedBagging	0.0293	0.0407	0.0114	37.2247	0.0000
	RandomUnderSampling	0.0212	0.0324	0.0112	27.1053	0.0000
	RandomOverSampler	0.0733	0.1612	0.0879	18.6969	0.0000
	SMOTE	0.0657	0.1569	0.0911	28.6228	0.0000
	SOMTEENN	0.0618	0.1535	0.0917	38.2040	0.0000
XGBoost	EasyEnsemble	0.0217	0.0463	0.0246	69.0403	0.0000
	BalancedBagging	0.0283	0.0556	0.0273	47.2147	0.0000
	RandomUnderSampling	0.0201	0.0437	0.0236	24.9083	0.0000
	RandomOverSampler	**0.0774**	**0.3092**	**0.2317**	**28.5667**	**0.0000**
	SMOTE	0.0434	0.2357	0.1923	47.8525	0.0000
	SOMTEENN	0.0444	0.2077	0.1634	37.2231	0.0000

Panel F. Forecasting period is the next six years

Model	Data Imbalanced Methods	FIN	FIN&SN	Difference	T_stat	p-value
RF	EasyEnsemble	0.0209	0.0318	0.0109	56.0079	0.0000
	BalancedBagging	0.0281	0.0402	0.0121	27.7606	0.0000
	RandomUnderSampling	0.0205	0.0325	0.0120	29.4787	0.0000
	RandomOverSampler	0.0712	0.1691	0.0978	17.1803	0.0000
	SMOTE	0.0658	0.1501	0.0842	21.3503	0.0000
	SOMTEENN	0.0648	0.1453	0.0806	21.8328	0.0000
XGBoost	EasyEnsemble	0.0211	0.0447	0.0236	58.4657	0.0000
	BalancedBagging	0.0270	0.0531	0.0261	37.4121	0.0000
	RandomUnderSampling	0.0197	0.0416	0.0219	21.3724	0.0000
	RandomOverSampler	**0.0882**	**0.3206**	**0.2324**	**23.7961**	**0.0000**
	SMOTE	0.0472	0.2218	0.1746	33.4958	0.0000
	SOMTEENN	0.0448	0.1980	0.1532	26.5554	0.0000

Note: FIN represents financial variables, and FIN&SN_C represents financial and social network variables.

Table 9. The Difference Tests Analyses in the Effectiveness of Bankruptcy Prediction Models: Before and After Introducing the Social Network Characteristics Variables (AUC)

Panel A. Forecasting period is the next one year

Model	Data Imbalanced Methods	FIN	FIN&SN_C	Difference	T_stat	p-value
RF	**EasyEnsemble**	**0.8274**	**0.8712**	**0.0438**	**7.8841**	0.0000
	BalancedBagging	0.8454	0.8282	-0.0172	-2.0463	0.0422
	RandomUnderSampling	**0.8124**	**0.8627**	**0.0503**	**8.4262**	0.0000
	RandomOverSampler	0.5172	0.5434	0.0261	8.5525	0.0000
	SMOTE	0.6227	0.6715	0.0488	5.4572	0.0000
	SOMTEENN	0.6631	0.6897	0.0266	2.5243	0.0125
XGBoost	**EasyEnsemble**	**0.8285**	**0.8688**	**0.0403**	**7.7118**	0.0000
	BalancedBagging	0.8163	0.8358	0.0195	3.2410	0.0014
	RandomUnderSampling	**0.7875**	**0.8610**	**0.0735**	**10.8247**	0.0000
	RandomOverSampler	0.5454	0.6461	0.1008	13.2211	0.0000
	SMOTE	0.6458	0.7071	0.0612	7.5173	0.0000
	SOMTEENN	0.6785	0.7112	0.0328	3.5941	0.0004

Panel B. Forecasting period is the next two years

Model	Data Imbalanced Methods	FIN	FIN&SN_C	Difference	T_stat	p-value
RF	**EasyEnsemble**	**0.8206**	**0.8332**	**0.0126**	**1.9213**	0.0566
	BalancedBagging	0.8427	0.7810	-0.0617	-8.7508	0.0000
	RandomUnderSampling	**0.8090**	**0.8250**	**0.0161**	**2.4490**	0.0155
	RandomOverSampler	0.5132	0.5366	0.0234	11.5505	0.0000
	SMOTE	0.6172	0.6419	0.0247	3.5892	0.0004
	SOMTEENN	0.6691	0.6638	-0.0053	-0.6307	0.5292
XGBoost	**EasyEnsemble**	**0.8192**	**0.8303**	**0.0111**	**2.2427**	0.0264
	BalancedBagging	0.8149	0.8063	-0.0085	-2.1230	0.0354
	RandomUnderSampling	**0.7821**	**0.8299**	**0.0478**	**8.9591**	0.0000
	RandomOverSampler	0.5397	0.6208	0.0811	15.6136	0.0000
	SMOTE	0.6472	0.6910	0.0438	7.4503	0.0000
	SOMTEENN	0.6838	0.7005	0.0166	1.9457	0.0536

Panel C. Forecasting period is the next three years

Model	Data Imbalanced Methods	FIN	FIN&SN_C	Difference	T_stat	p-value
RF	EasyEnsemble	0.8228	0.7907	-0.0321	-6.5677	0.0000
	BalancedBagging	0.8493	0.7531	-0.0962	-23.6328	0.0000
	RandomUnderSampling	0.8094	0.7946	-0.0148	-2.8386	0.0053
	RandomOverSampler	0.5126	0.5390	0.0265	14.7190	0.0000
	SMOTE	0.6062	0.6176	0.0114	2.0812	0.0396
	SOMTEENN	0.6723	0.6446	-0.0277	-4.2772	0.0000
XGBoost	EasyEnsemble	0.8144	0.8014	-0.0130	-3.9883	0.0001
	BalancedBagging	0.8172	0.7912	-0.0260	-7.8492	0.0000
	RandomUnderSampling	**0.7745**	**0.8091**	**0.0346**	**6.8863**	0.0000
	RandomOverSampler	0.5382	0.6184	0.0803	18.3784	0.0000
	SMOTE	0.6444	0.6721	0.0277	5.8232	0.0000
	SOMTEENN	0.6890	0.6820	-0.0070	-0.9841	0.3271

Panel D. Forecasting period is the next four years

Model	Data Imbalanced Methods	FIN	FIN&SN_C	Difference	T_stat	p-value
RF	EasyEnsemble	0.8235	0.7735	-0.0500	-17.2842	0.0000
	BalancedBagging	0.8477	0.7487	-0.0990	-32.1700	0.0000
	RandomUnderSampling	0.8115	0.7781	-0.0334	-8.2182	0.0000
	RandomOverSampler	0.5143	0.5448	0.0305	22.8565	0.0000
	SMOTE	0.6211	0.6101	-0.0110	-2.8015	0.0062
	SOMTEENN	0.6910	0.6368	-0.0542	-11.4834	0.0000
XGBoost	EasyEnsemble	0.8092	0.8013	-0.0079	-2.3070	0.0234
	BalancedBagging	0.8118	0.7860	-0.0259	-7.1202	0.0000
	RandomUnderSampling	**0.7724**	**0.8040**	**0.0315**	**6.4201**	0.0000
	RandomOverSampler	0.5401	0.6233	0.0831	23.4705	0.0000
	SMOTE	0.6472	0.6679	0.0207	4.0771	0.0001

	SOMTEENN	0.7025	0.6768	-0.0258	-4.1057	0.0001
Panel E. Forecasting period is the next five years						
Model	Data Imbalanced Methods	FIN	FIN&SN_C	Difference	T_stat	p-value
RF	EasyEnsemble	0.8187	0.7677	-0.0510	-22.7143	0.0000
	BalancedBagging	0.8463	0.7649	-0.0814	-27.8414	0.0000
	RandomUnderSampling	0.8082	0.7768	-0.0314	-7.0663	0.0000
	RandomOverSampler	0.5193	0.5445	0.0252	17.9145	0.0000
	SMOTE	0.6264	0.6085	-0.0179	-4.8778	0.0000
	SOMTEENN	0.6956	0.6372	-0.0584	-11.2109	0.0000
XGBoost	EasyEnsemble	0.8048	0.7914	-0.0134	-4.1461	0.0001
	BalancedBagging	0.8173	0.7835	-0.0338	-9.8235	0.0000
	RandomUnderSampling	**0.7767**	**0.8043**	**0.0276**	**4.6226**	0.0000
	RandomOverSampler	0.5481	0.6160	0.0679	17.2898	0.0000
	SMOTE	0.6581	0.6546	-0.0034	-0.7414	0.4614
	SOMTEENN	0.7113	0.6620	-0.0493	-10.0017	0.0000
Panel F. Forecasting period is the next six years						
Model	Data Imbalanced Methods	FIN	FIN&SN_C	Difference	T_stat	p-value
RF	EasyEnsemble	0.8121	0.7824	-0.0296	-13.6739	0.0000
	BalancedBagging	0.8376	0.7731	-0.0645	-14.5939	0.0000
	RandomUnderSampling	0.7996	0.7868	-0.0128	-2.8106	0.0088
	RandomOverSampler	0.5189	0.5472	0.0282	16.6727	0.0000
	SMOTE	0.6299	0.6044	-0.0255	-5.4296	0.0000
	SOMTEENN	0.7120	0.6311	-0.0809	-17.3023	0.0000
XGBoost	**EasyEnsemble**	**0.8000**	**0.8001**	**0.0001**	**0.0294**	0.9767
	BalancedBagging	0.8082	0.7900	-0.0182	-3.8396	0.0006
	RandomUnderSampling	**0.7764**	**0.8112**	**0.0348**	**4.5396**	0.0001
	RandomOverSampler	0.5581	0.6285	0.0705	12.4572	0.0000
	SMOTE	0.6764	0.6595	-0.0169	-2.8244	0.0085
	SOMTEENN	0.7194	0.6643	-0.0552	-8.5823	0.0000

Note: FIN represents financial variables, and FIN&SN_C represents financial and social network variables.

Table 10: Ranking of Importance of Managers' Social Network Characteristics and Financial Variables (Each Training Period)

Panel A. Training period: 2000~2014

Variable	Feature Importance	Variable	Feature Importance
(1) EMV_Debt	0.062568	(8) ROE_Chg	0.038225
(2) RE_TA	0.046091	**(9) std_Degree_nCEO**	**0.037050**
(3) NWC_TA	0.044686	**(10) Degree_all**	**0.036410**
(4) Betweenness_CEO	**0.043930**	(11) PB_Chg	0.036220
(5) EBIT_TA	0.041139	**(12) Degree_CEO**	**0.036151**
(6) Sales_TA	0.040208	**(13) Degree_nCEO**	**0.035896**
(7) EBIT_Sales	0.039920	**(14) std_Eigenvector_nCEO**	**0.035788**

Panel B. Training period: 2000~2015

Variable	Feature Importance	Variable	Feature Importance
(1) EMV_Debt	0.061963	(8) TA_growth	0.037996
(2) RE_TA	0.045756	(9) Sales_TA	0.037775
(3) NWC_TA	0.042205	**(10) Degree_CEO**	**0.037268**
(4) Betweenness_CEO	**0.041895**	**(11) Degree_nCEO**	**0.036983**
(5) EBIT_TA	0.039874	(12) EMP_growth	0.036878
(6) std_Degree_nCEO	**0.039006**	(13) EBIT_Sales	0.036772
(7) ROE_Chg	0.038986	**(14) std_Degree_all**	**0.036485**

Panel C. Training period: 2000~2016

Variable	Feature Importance	Variable	Feature Importance
(1) EMV_Debt	0.063094	(8) EBIT_TA	0.038467
(2) NWC_TA	0.046432	**(9) std_Degree_all**	**0.038396**
(3) RE_TA	0.042805	**(10) Degree_all**	**0.038151**
(4) Betweenness_CEO	**0.041925**	**(11) Degree_nCEO**	**0.037995**
(5) Sales_TA	0.041528	(12) EBIT_Sales	0.037466
(6) std_Degree_nCEO	**0.040235**	**(13) Degree_CEO**	**0.037384**
(7) ROE_Chg	0.038890	(14) TA_growth	0.036546

Panel D. Training period: 2000~2017

Variable	Feature Importance	Variable	Feature Importance
(1) EMV_Debt	0.063088	**(8) Degree_all**	**0.037540**
(2) RE_TA	0.046443	**(9) std_Degree_all**	**0.037485**
(3) NWC_TA	0.045693	(10) TA_growth	0.037146
(4) Betweenness_CEO	**0.042072**	(11) ROE_Chg	0.037020
(5) std_Degree_nCEO	**0.040052**	(12) EBIT_Sales	0.036990
(6) EBIT_TA	0.039541	(13) Sales_TA	0.036984
(7) Degree_nCEO	**0.038320**	**(14) Degree_CEO**	**0.036799**

Panel E. Training period: 2000~2018

Variable	Feature Importance	Variable	Feature Importance
(1) EMV_Debt	0.061382	(8) ROE_Chg	0.039691
(2) RE_TA	0.045721	**(9) Degree_nCEO**	**0.039655**
(3) NWC_TA	0.04451	**(10) Degree_CEO**	**0.038611**
(4) Betweenness_CEO	**0.042576**	(11) Sales_TA	0.038528

(5) EBIT_Sales	0.040953	(12) TA_growth	0.037095
(6) std_Degree_nCEO	**0.040563**	**(13) Degree_all**	**0.037087**
(7) std_Degree_all	**0.039738**	(14) EBIT_TA	0.036925

Panel F. Training period: 2000~2019

Variable	Feature Importance	Variable	Feature Importance
(1) EMV_Debt	0.062223	(8) ROE_Chg	0.038683
(2) RE_TA	0.045412	**(9) Degree_nCEO**	**0.037095**
(3) NWC_TA	0.044334	**(10) Degree_all**	**0.036912**
(4) Betweenness_CEO	**0.041302**	(11) EBIT_Sales	0.036507
(5) std_Degree_nCEO	**0.040773**	(12) EBIT_TA	0.036406
(6) std_Degree_all	**0.040227**	(13) EMP_growth	0.036205
(7) TA_growth	0.038933	**(14) Degree_CEO**	**0.035166**