Distressed Chinese firm prediction with discretized data

Jun Huang
A.R. Sanchez, Jr. School of Business,
Texas A&M International University, Laredo, Texas, USA
Haibo Wang
Division of International Business & Technology Studies,
Texas A&M International University, Laredo, Texas, USA, and
Gary Kochenberger
Department of Business Analytics, University of Colorado at Denver,
Denver, Colorado, USA

Received 6 August 2016
Revised 4 March 2017
Accepted 21 March 2017

Abstract

Purpose – The authors develop a framework to build an early warning mechanism in detecting financial deterioration of Chinese companies. Many studies in the financial distress and bankruptcy prediction literature rarely do they examine the impact of pre-processing financial indicators on the prediction performance. The purpose of this paper is to address this shortcoming.

Design/methodology/approach – The proposed framework is evaluated by using both original and discretized data, and a least absolute shrinkage and selection operator (LASSO) selection technique for choosing an appropriate subset of financial ratios for improved predictive performance. The financial ratios are then analyzed by five different data mining techniques. Managerial insights, using data from Chinese companies, are revealed by the methodology employed.

Findings – The prediction accuracy increases after we discretized the continuous variables of financial ratios. A better prediction performance can be achieved by including fewer, but relatively more significant variables. Random forest has the highest overall performance following closely by SVM and neural network.

Originality/value – The contribution of this study is fourfold. First, the authors add to the literature on defaults by showing variable discretization to be an essential pre-processing step to improve the prediction performance for classification problems. Second, the authors demonstrate that machine learning approaches can achieve better performance than traditional statistical methods in classification tasks. Third, the authors provide the evidence for the adoption of C5.0 over other methods because rules generated with C5.0 provide managerial insights for managers. Finally, the authors demonstrate the effectiveness of the LASSO technique for identifying the most important financial ratios from each category, enabling one to build better predictive models.

Keywords LASSO, Data mining, Support vector machine, Financial ratios, Distress, Random forests

Paper type Research paper

Introduction

In the past three decades, China banking systems have been reformed to meet the demands of the tremendous economic growth that has been experienced. During this period of the high growth rate of GDP and large volume of FDI to boost China's economy, the banking systems have employed many risk prevention measures on business/personal loans. The Chinese Government responded to the global financial crisis in 2008 with huge investments in infrastructure and kept the real estate market floating to avoid the exposure of non-performing loans (NPLs) in the banking systems even though it had reached a crisis level (Suzuki *et al.*, 2008). However, the recent slowdown in the growth rate in China and the return of inefficient infrastructure investment have brought new attention to be placed on NPLs. A recent report by Reuters stated that NPLs more than doubled in 2015 from 2014 (Lian, 2016) and some investment firms and managers warned their investors about the credit risk on China banks (Porzecanski, 2016; Osborn *et al.*, 2015). In the meantime, additional research papers have been published on NPLs of Chinese banking systems in the past few

Management Decision
Vol. 55 No. 5, 2017
pp. 786-807

0025-1747
DOI 10.1108/MD-08-2016-0546

years compared to the last decade (Griffiths, 2005; Lu *et al.*, 2005; Suzuki *et al.*, 2008; Potena, 2013; Gan *et al.*, 2014; Cai and Huang, 2014; Zhu *et al.*, 2015; Zha *et al.*, 2016; Zhang *et al.*, 2016).

To mitigate the risk of NPLs in the banking systems, there are many measures in place such as the firm-level credit risk prediction, which can be used to evaluate business loan applications. Firm-level credit risk prediction, such as financial distress prediction, bankruptcy prediction, and default risk prediction, has been a popular and interesting topic for decades due to its importance to bankers, investors, and firms, alike. Being able to reliably forecast the financial distress of firms and financial institutions can lower the level of NPLs, enabling investors to adjust their investment strategies to reduce losses, and firm CEOs can establish a warning mechanism for financial deterioration in an early stage (Lacher *et al.*, 1995; Geng *et al.*, 2015). After the seminal study from Altman (1968), adopting discriminant analysis for corporate bankruptcy prediction based on a number of financial ratios, a great number of studies attempted to predict company financial distress or bankruptcy with the financial ratios using different statistical and data mining techniques (McKee and Lensberg, 2002; Leshno and Spector, 1996; Lee and Chen, 2005; Shin *et al.*, 2005; Yang *et al.*, 2011; Chen and Du, 2009; Gestel *et al.*, 2006; Fedorova *et al.*, 2013; Li *et al.*, 2014; Sun *et al.*, 2014; Geng *et al.*, 2015; Liang *et al.*, 2016).

In this study, we propose a predictive analytics framework using a set of data mining techniques, namely, C5.0, support vector machines (SVMs), random forests (RFs), neural networks (NN), linear discriminant analysis (LDA) and logistic regression (LR), to identify the effective techniques for predicting financial distress for Chinese companies. Most previous studies focus on applying state-of-the-art data mining techniques in achieving better prediction performance, but rarely have examined the impact of pre-processing of the financial ratios on the prediction performance. Therefore, an important objective of this study is to test whether discretization of continuous variables, mainly financial ratios, can improve the prediction performance of this classification problem. Our motivation comes from the literature showing that discretization of continuous data in classification problems can significantly impact the performance of classification algorithms, especially for machine learning algorithms such as SVM and decision trees (DTs) (Lustgarten *et al.*, 2008; Bolon-Canedo *et al.*, 2009; Tillander, 2012).

Another objective of this study is to identify fewer, but relatively more important financial ratios in the prediction model, and to see if they can provide better prediction performance in terms of classification accuracy. On one hand, being practical, we do not want to keep too many variables which will increase the complexity of the model and make it harder to interpret. On the other hand, evaluation with selected features may improve the prediction accuracy (Guyon and Elisseeff, 2003; Tian *et al.*, 2015; Miller, 1984). Therefore, we apply the least absolute shrinkage and selection operator (LASSO) technique for selecting financial ratios. We further explore the top ranked ratios from each financial category and build parsimonious models based on these ratios to increase interpretability of a model and provide managerial insights in detecting financial distress of Chinese companies. The framework in this study is based on widely used data mining techniques for predicting financial distress, but these techniques are rarely used to treat data from public listed Chinese companies.

The rest of the paper is organized as follows. The second section briefly reviews literature in prediction of financial distress using various feature selection and data mining techniques. The third section introduces the framework in this study. The fourth section describes the test data. The fifth section shows the results. Finally, conclusion is given in the sixth section.

Literature review

There is rich literature on predicting corporate financial distress with empirical evidence or empirical analysis models in the last five decades. Keasey and Watson (1991) provided a review of financial distress prediction models in the literature and discuss the direction of research in the field. Demiroglu and James (2011) presented a comprehensive review on the use of bank lines of credit as a source of corporate liquidity based on empirical evidence. Altman (1968) first proposed to use discriminant analysis on financial ratios as empirical evidence. Financial ratios have been found to be useful empirical evidence in many studies as we summarize in Table I. DeAngelo and DeAngelo (1990) used dividend reduction as evidence for predicting financial distress but also pointed out its use as a strategy for bargaining with organized labor. Hoshi *et al.* (1990) discussed the role of the banking system in Japan to reduce the cost of financial distress. John (1993) reported a study analyzing the relationship of the costs of financial distress to the level of corporate liquidity maintained and leverage using linear models. Altman *et al.* (1994) compared the performance of LDA and NN on distress classification and prediction and pointed out the potential problem of overfitting NN models. Bhagat *et al.* (1994) studied financial distress using law suit data as evidence. Daily and Dalton (1994) examined the relationship between firm financial distress and its governance structures using logistics regression model. Opler and Titman (1994) analyzed the indirect costs of financial distress of highly leveraged firms, highlighting that the more conservatively financed competitors will survive industry downturns. Alderson and Betker (1995) used empirical data from the firms under chapter 11 and pointed out that the choice of capital structures by the firms is determined by the liquidation costs of assets. Theodossiou *et al.* (1996) used empirical data to examine the economic factors played in the acquisition of financial distressed firms and assets. Sudarsanam and Lai (2001) evaluated the effectiveness of turnaround strategies using financial ratios from recovery and non-recovery firms. Platt and Platt (2002) discussed choice-based sample bias when researchers applied financial ratios as empirical evidence and argued that all firms should be included in the population to build an effective early warning model. Almeida *et al.* (2011) examined a model for investment policies using empirical data and pointed out several new predictions contradictory to the literature because they have never been empirically examined. Almamy *et al.* (2016) evaluated the extension of the Altman's *Z*-score model by adding a new variable and showed that cash flow is highly significant in predicting the health of UK companies in terms of predictive power. Li *et al.* (2014c) used standard financial ratios and corporate efficiency to predict corporate distress in Chinese companies. They found that the predictive power of the model is improved by using corporate efficiency information which was measured with data envelopment analysis (DEA). Geng *et al.* (2015) predicted financial distress of Chinese companies with various data mining techniques and found that NNs performed better than other classifiers. Table I reports the financial ratios in some of the aforementioned studies.

This study is related to a large body of work on data mining techniques for predicting corporate financial distress, namely, DA, LR, SVM, NN, RF, and C5.0 (Kumar and Ravi, 2007; Sinha and Zhao, 2008; Kwak *et al.*, 2012; Olson *et al.*, 2012; Korol, 2013; Tsai and Hsu, 2013). Discriminant analysis and LR analysis are the most frequently used statistical techniques in predicting business failure. DA was first adopted by Altman (1968) in predicting corporate bankruptcy. Following his lead, Lawrence and Bear (1986) employed discriminant analysis on a data set consisting of 42 bankrupt firms and 42 non-bankrupt firms for the period 1975-1981 and reported that capitalization of leases did not significantly improve the classification accuracy of the bankruptcy models. To overcome some limitations of DA due to its restricted assumptions, researchers proposed LR analysis because of its nature in providing binary prediction results. Zavgren *et al.* (1988) applied LR to examine the association between model-derived probabilities of failure and market reactions to the news of company financial distress. Senteney *et al.* (2006) reported that the log-linear LR model can provide explanatory power of auditor-qualified opinions and traditional financial statement ratios in prediction of impending bankruptcy. Youn and Gu (2010) compared the performance of LR and artificial neural networks (ANNs) for predicting financial distress of

Table I. Financial ratios in financial distress and bankruptcy prediction literatures

Reference	No. of financial ratios	Names of financial ratios used
Altman (1968)	5	Working capital/total assets; retained earnings/total assets; EBIT/total assets; market value equity/book value of debt; sales/total assets
Frydman *et al.* (1985)	20	Cash/total assets; cash/total sales; cash flow/total debt; current assets/current liabilities; current assets/total assets; current assets/total sales; EBIT/total assets; log (interest coverage + 15); log (total assets); market value of equity/total capitalization; net income/total assets; quick assets/total assets; quick assets/current liabilities; quick assets/sales; retained earnings/total assets; standard deviation of (EBIT/total assets); total debt/total assets; total sales/total assets; working capital/total assets; working capital/total sales
Leshno and Spector (1996)	70	Working capital/total sales; retained earnings/total assets; earnings before income tax/total assets; market value/total liabilities; sales/total assets; EBIT per share; cash flow per share; cost of goods sold/sales; capital expenditures per share; sales/cash; receivables turnover; inventory turnover; ROE; ROI; investments/assets; long-term debt/total liabilities; debt/equity; long-term debt/equity; quick ratio; price/earnings ratio; dividend yield; total debt/total assets; quick assets/sales; sales/total capital; log (total assets); interest coverage; log (interest coverage); earning/5 years maturity; cash flow/total debt; working capital/long-term debt; working capital/cash expenses; book equity/total capital; market equity/total capital; average market equity/total capital; StDv (EBIT/total assets); StDv (log (EBIT/total assets)); sales/gross fixed assets; sales/receivables; ROA; total debt/invested capital; current ratio; worth/total debt; net income/total debt; operating income/sales; EBIT/total tangible assets; net available for capital/total capital; sales/total tangible assets; EBIT/sales; current liabilities/total liabilities; net available for total capital/sales; fixed charge coverage; cash flow/fixed charges; earning/total debt; retaining earning/tangible assets; capital lease/total assets; EBIT drop; average short-term borrow; number years of negative profit; sales per share; net profit margin; cash flow margin; fixed charge coverage; margin drop; auditor; auditor opinion; number of employees; pension expenses; bond rating; total investment
McKee and Lensberg (2002)	9	General and administration expense/net sales; net income/net worth; current assets/current liabilities; liabilities/total assets; net worth/net fixed assets; working capital/net worth; net income/total assets; cash/current liabilities; investment cash flow/net income
Ryu and Yue (2005)	23	Cash flow/total assets; cash/sales; cash flow/total debt; current assets/current liabilities; current assets/total assets; current assets/sales; EBIT/total assets; retained earnings/total assets; net income/total assets; total dent/total assets; sales/total assets; working capital/total assets; working capital/sales; quick assets/total assets; quick assets/current liabilities; quick assets/sales; market value of equity/total capitalization; cash/current liabilities; current liabilities/equity; inventory/sales; equity/sales; market value of equity/total debt; net income/total capitalization
Shin *et al.* (2005)	10	Total asset growth; contribution margin; operating income to total asset; fixed asset to sales; owner's equity to total asset; net asset to total asset; net loan dependence rate; operating asset constitute ratio; working capital turnover period, net operating asset turnover period
Etemadi *et al.* (2009)	43	EBIT/total assets; long-term debt/shareholders' equity; retained earnings/stock capital; market value of equity/total liabilities; market value equity/shareholders' equity; market value equity/total assets; cash/

(*continued*)

Reference	No. of financial ratios	Names of financial ratios used
		total assets; size(log total asset); total liabilities/total assets; current liabilities/shareholders' equity; current liabilities/total liabilities; (cash + short-term investments)/current liabilities; (receivables + inventory)/total assets; receivables/sales; receivables/inventory; shareholders' equity/total liabilities; shareholders' equity/total assets; current assets/current liabilities; quick assets/current liabilities; quick assets/total assets; fixed assets/(shareholders' equity + long-term debt); fixed assets/total assets; current assets/total assets; cash/current liabilities; interest expenses/gross profit; sales/cash; sales/total assets; working capital/total assets; paid in capital/shareholders' equity; sales/working capital; retained earnings/total assets; net income/shareholders' equity; net income/sales; net income/total assets; operational income/sales; operational income/total assets; EBIT/interest expenses; EBIT/sales; gross profit/sales; sales/shareholders' equity; sales/fixed assets; sales/current assets
Min and Jeong (2009)	27	Gross value added/sales; gross value added/total assets; growth rate of total assets; ordinary income/sales; net income/sales; operating income/sales; costs of sales/sales; net interest expenses/sales; ordinary income/total assets; rate of earnings on total capital; net working capital/total assets; current liabilities/total assets; stockholders' equity/total assets; total borrowings and bonds payable/total assets; total assets turnover; ordinary income/total assets; net working capital/sales; stockholders' equity/sales; ordinary income/total assets; depreciation expenses; operating assets turnover; interest expenses/total expenses; net interest expenses; break-even point ratio; employment costs; interest expenses and net income/total assets; earnings before interest and taxes/sales
Fedorova *et al.* (2013)	83	Cash flow/total liabilities; cash flow/equity; cash flow/total sales; cash flow/total assets; cash flow/equity; cash flow/current liabilities; cash flow/total assets; cash flow/total sales; cash flow/current liabilities; gross profit/total sales; gross profit/total assets; EBT/total liabilities; profit on sales/total sales; profit on sales/total assets; net income/total liabilities; EBT/total sales; EBT/total assets; profit on sales/current liabilities; gross profit/cost of goods sold; profit on sales/equity; net profit/current liabilities; profit on sales/cost of goods sold; gross profit/total liabilities; gross profit/current liabilities; EBT/cost of goods sold; gross profit/equity; profit on sales/total liabilities; net profit/cost of goods sold; sales/fixed assets; sales/equity; (cost of goods sold – depreciation)/accounts payable; sales/current assets; sales/total liabilities; (cost of goods sold – depreciation)/inventories; sales/(cash + invested funds); sales/current liabilities; sales/(cash + invested funds + accounts receivable); sales/accounts receivable; sales/working capital; cost of goods sold/finished goods; cash/current liabilities; short-term accounts receivable/accounts payable; (cash + invested funds)/(costs/365); (equity-fixed assets)/current assets; quick assets/(costs/365); quick assets/total assets; long-term liabilities/equity; cash/total assets; quick assets/current assets; current assets/total liabilities; cash/current assets; short-term liabilities/total liabilities; current assets/total assets; revenue reserves/equity; long-term liabilities/fixed assets; (cash + invested funds)/total assets; revenue reserves/total assets; long-term liabilities/total liabilities; (equity + long-term liabilities)/total assets; revenue reserves/total liabilities; current liabilities/total liabilities; working capital/inventories; long-term

Table I. (*continued*)

Reference	No. of financial ratios	Names of financial ratios used
		liabilities/total assets; accounts payable/total liabilities; retained earnings/equity; fixed assets/total assets; accounts payable/accounts receivable; log (tangible total assets); debt/total assets; profit before tax/current liabilities; working capital/total debt; equity/total liabilities; working capital/total assets; log (EBIT)/interest net profit/costs; retained earnings/total assets; EBT/equity current liabilities/(cash + invested funds); sales/total assets; EBIT/total assets; total assets/sales; cash flow/total debt; no-credit interval; current liabilities/total assets; net profit/equity
Li *et al.* (2014)	35	Operating revenue per share; return on equity (ROE); return on assets (ROA); return on invested capital (ROIC); gross margin/total sales; operating profit/total sales; operating expenses/total sales; financial expenses/total sales; undistributed profits per share; EBIT per share (EBITPS); current liabilities/total liabilities; current ratio; quick ratio; cash ratio; EBITDA/total liabilities; surplus capital per share; surplus reserve per share; book value per share (BPS); equity multiplier; current assets/total assets; tangible assets/total assets; net cash flow from operating per share; net cash flow per share; net cash flow from operating/operating revenue; net cash flow from operating/total liabilities; net cash flow from operating/interest bearing liabilities; net cash flow from operating/current liabilities; inventory turnover; receivables turnover; current assets turnover; operating revenue growth; total profit growth; net profit growth; total assets growth
Geng *et al.* (2015)	31	Total liabilities/total assets; current assets/current liabilities; (current assets-inventory)/current liabilities; total liabilities/total shareholders' equity; current liabilities/total assets; net operating cash flow/current liabilities; earnings before interest and tax (EBIT)/interest expense; (sales revenue-sales cost)/sales revenue; net profit/sales revenue; earnings before income tax/average total assets; net profit/average total assets; net profit/average current assets; net profit/average fixed assets; net profit/average shareholders' equity; business income/average total assets; sales revenue/average current assets; sales revenue/average fixed assets; main business cost/average inventory; main business income/average balance of accounts receivable; cost of sales/average payable accounts; main business income of this year/main business income of last year; total assets of this year/total assets of last year; net profit of this year/net profit of last year; current assets total assets; fixed assets/total assets; shareholders' equity/fixed assets; current liabilities/total liabilities; net profit/number of ordinary shares at the end of year; net assets/number of ordinary shares at the end of year; net increase in cash and cash equivalents/number of ordinary shares at the end of year; capital reserves/number of ordinary shares at the end of year

Table I.

US restaurant firms where LR model not only outperformed ANN, but also guided the firm to the factors of bankruptcy risk. Foster and Zurada (2013) applied LR as a feature selection method and constructed an adjustable hazard model to improve the predictive accuracy for financially distressed samples. Hilston Keener (2013) adopted a LR model to study the financial distress of retail industrial and found a few financial ratios linked to the bankruptcy such as lower cash to current liability ratios, lower cash flow margins, and higher debt to equity ratios. Li *et al.* (2014) reported the use of LR and DEA for predicting corporate distress in Chinese companies and showed how the corporate efficiency information provided by DEA model can improve the prediction accuracy of LR.

Besides statistical techniques, machine learning methods based on artificial intelligence have become dominant methods for solving such classification problems. SVM, NN, and DTs were among the most commonly used machine learning techniques. Bellotti and Crook (2009) tested SVMs against traditional methods, LR, and discriminant analysis, on a large credit card database. They found that SVMs perform competitively well, and unlike many other learning tasks, a large number of support vectors are required to achieve the best performance due to the nature of the credit data for which the available application data can only be broadly indicative of default. Shin *et al.* (2005) evaluated the predictive performance of bankruptcy based on the selected ratios with SVMs. Compared with back-propagation neural network (BPN), they found that generalization performance of SVM is better than that of BPN, as the training set size reduces. Although many studies applied SVM in prediction models, Tsai (2008) pointed out that the performance of SVMs is not fully understood in the literature because an insufficient number of data sets have been considered and different kernel functions are used to train the SVMs. Härdle *et al.* (2009) reported on exploring the suitability of smooth support vector machines to examine the important factors on influencing the precision of prediction. Dellepiane *et al.* (2015) propose new country-specific factors using SVM as the forecasting model and assess the general effectiveness of SVMs by comparing it with the performances of other commonly used methods.

NNs represent a popular data mining technique in financial prediction due to its "blackbox" feature of handling different types of information with high flexibility. Wuerges and Borba (2010) reported that ANN is the most popular methods in the literature when they reviewed the published research works from 2000 to 2007 on challenged problems in Finance and Accounting. Lee *et al.* (1996) developed the hybrid NN models and evaluated its performance using Korean bankruptcy data with promising results in terms of predictive accuracy and adaptability. Jain and Nag (1997) discussed the critical issues affecting the performance of NNs including training sample design and the use of an appropriate performance metric. Luther (1998) reported a study on the data set of 104 firms that filed for bankruptcy under chapter 11 using an NN model trained by the genetic algorithm to avoid the local minima. Yang *et al.* (1999) pointed out in their study that probabilistic NNs without pattern normalization and Fisher discriminant analysis achieve the best overall estimation results. Zhang *et al.* (1999) presented a general framework using ANNs in bankruptcy prediction. Their results indicated that ANN-based models are significantly better than LR models in prediction as well as classification rate estimation in addition to strength.

DTs are another popular approach for addressing classification problems. Olson *et al.* (2012) illustrated their preference for DT to predict corporate failure. They argued that DT could provide models with transparency, transportability, and accuracy. RF and C5.0 are two relatively new DT techniques with considerable promise. Whiting *et al.* (2012) reported that ensemble methods in machine learning such as RF shows practical potential in terms of accuracy and interpretability. Fernndez-Delgado *et al.* (2014) evaluated 179 classifiers with 121 data sets, finding RF to be the best classifier. There are limited studies on the classification tree approach C5.0, an improved version of C4.5. C4.5 has been used as the benchmark for ensemble methods, and it can achieve acceptable results on the small data sets, but lagged behind other advanced techniques such as ANN and Memetic Algorithm (Pendharkar, 2005; Karami *et al.*, 2012). Finally, there are a great number of studies comparing data mining techniques. Kumar and Ravi (2007) presented a comprehensive review of this research between 1968 and 2005. Sinha and Zhao (2008) published a study comparing the performance of seven data mining classification methods – naive bayes, LR, DT, decision table, NN, *K*-nearest neighbor, and SVM – with and without incorporating domain knowledge. Kwak *et al.* (2012) evaluated the data mining applications on Korean bankruptcy data after the 1997 financial crisis and proposed a multiple-criteria linear programming method to improve the prediction accuracy. Olson *et al.* (2012) applied a variety of data mining tools in their study and found DTs to be relatively more accurate compared to NNs and SVMs in their data set. Korol (2013) reported a study comparing the effectiveness of discriminant analysis, decisional trees, and ANN models using data sets from Latin America and Central Europe. Tsai and Hsu (2013) proposed a meta-learning framework, which is composed of two-level classifiers for bankruptcy prediction. The results of their study show that the proposed framework outperformed the basic techniques of NNs, DTs, and LR methods alone. Based on the finding of these papers, we develop a framework of using a set of data mining techniques and LASSO on selecting discretized data.

The analytical framework

In this study, predictive analytics models are built using four machine learning techniques, namely, C5.0, RF, SVMs, and NN, and two traditional statistical techniques, namely, LDA, and LR in order to compare prediction performance of these data mining techniques. Unlike many previous studies which typically randomly divide the data into a training set and a testing set with certain partition ratio, this study divides data sample in a chronological sequence. The models are trained and cross-validated on data from 2003 to 2009 and tested on an-out-of-time test set from 2010 to 2011. A total of 95 financial indicators (ratios) are considered.

This study also examines whether discretizing continuous data in a classification problem can improve the classification performance. We first test the models with data in its original continuous form. We then discretize the data for all the ratios with a quantile-based discretization function which discretize variables into equal-sized buckets based on sample quantiles. Finally, besides evaluating the prediction performance with all 95 variables, we examine the prediction performance based on a subset of variables selected by the LASSO technique.

Data collection and preparation

The data sample was derived from the China Security Market Accounting Research (CSMAR) database provided by GTA, a leading global provider of China financial market, industries, and economic data. The database also provides financial ratios grouped in seven categories, namely, cash flow indicators, profitability indicators, liquidity indicators, solvency indicators, shareholders' profitability indicators, operating indicators, and leverage indicators. All the companies represented are from the manufacturing sector. After discarding the ratios with more than 30 percent missing values, we keep 95 financial ratios in this study. The financial ratios' code and formulae for these ratios are given in Table AI. The missing values in these 95 ratios are imputed with the mean for the corresponding company.

In addition, we randomly selected 156 non-ST companies to match the number of ST companies in order to avoid unbalanced sample sizes between the two classes. The companies labeled ST are considered financial distressed companies, and are denoted with the value of one while non-ST companies are denoted with the value of 0. Unlike previous studies on Chinese ST companies that focused on one or two years ahead of ST, this study uses financial data three years prior to ST to predict financial distress of a company. According to the disclosure policy of Chinese listing companies, the announcement for a company to be ST at year t is mainly based on its financial performance in the past two years, and thus using financial ratios from year $t-1$ or $t-2$ to predict the ST status at year t will raise the problem of overestimating the predictive power of a model. Therefore, we try to predict the ST status of a company with financial data from year $t-3$ to examine the performance of the models. The data for the label variable, namely, ST or non-ST is from 2003 to 2011, but the corresponding financial ratios data are three years earlier from 2000 to 2008.

Selection of important financial ratios by LASSO

As shown in Table I, the research community has adopted as many as 83 variables for use in financial distress prediction modeling. This large number of variables can increase variable collinearity and lead to greater variance in the predictive model performance. In addition, including irrelevant and redundant variables can lead to a poor predictive accuracy due to high complexity, intensive computation, and instability. Therefore, many studies suggest using a subset of variables from a number of candidate financial ratios using various selection methods such as independent samples *t*-test, ANOVA test, discriminant analysis, sequential elimination, mutual information-based feature selection, etc. (Ryu and Yue, 2005; Shin *et al.*, 2005; Etemadi *et al.*, 2009; Min and Jeong, 2009; Fedorova *et al.*, 2013).

In this study, we use the LASSO technique to rank the importance of all the 95 financial ratios, and select the top ratios from each financial category based on the ranking. The LASSO technique was proposed by Tibshirani (1996) and has since gained popularity for its success in both feature selection and ridge regression. The idea is to impose a limit on the sum of absolute values of the regression coefficients, enabling some coefficients to go to 0, exposing insignificant variables. The LASSO model can be described as follows.

Given a set of independent variables $x_1, x_2, \ldots, x_n$ and a dependent variable y, the OLS estimator for dependent variable:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

and the LASSO function can be defined as:

$$\text{Min} \sum (y - \hat{y})^2$$

s.t.:

$$\sum |\beta_i| \leqslant s \text{ where } i = 1, \ldots, n$$

By decreasing the value of s, some of β_i are forced to be 0, effectively removing the variables from the model.

The advantages of LASSO over some traditional feature selection methods, such as stepwise selection, include its consistency in light of small perturbations of data changes and its tendency to naturally overcome the multi-collinearity problem (Tian *et al.*, 2015). The LASSO estimation, as a function of the shrinkage, illustrates the order in which variables enter the model as one relaxes the constraint on the L1 norm of their estimates. Therefore, it provides an entire variable selection path. Many studies in recent years have highlighted the high level of success of LASSO for variable selection. Lustgarten *et al.* (2008) proposed a procedure to combine Transductive LASSO and Dantzig Selector for prediction of high-dimensional problems and Bolon-Canedo *et al.* (2009) reported a similar finding on LASSO and the Dantzig selector for high-dimensional regression with noise. West (2000) presented a study to extend the adaptive LASSO (ALASSO) approach for variable selection and report favorable results on data from the US Department of Agriculture's Continuing Survey. Kaul (2014) also reported a simulation study to analyze the performance of adaptive LASSO. Lacher *et al.* (1995) proposed a LASSO procedure for estimating a threshold autoregressive model and applied it to the quarterly US real GNP data from 1947 to 2009. Tian *et al.* (2015) applied LASSO as a variable selection procedure to a comprehensive bankruptcy database and reported that LASSO outperformed other variable selection models. Guyon and Elisseeff (2003) presented a new Bayesian LASSO to select the influential parameters. Lee and Chen (2005) discussed a study using LASSO and a ridge regression approach to develop empirical models for bankruptcy prediction and applied their approach to a data set from the hospitality industry. Efron *et al.* (2004) reported a study on bootstrap Granger Causality test using an adaptive LASSO procedure on high-dimensional forecasting problem of macro-economic developments and compared it with the standard Wald test. Gestel *et al.* (2006) proposed a new LASSO regression model and suggest that it outperforms many feature selection methods for handling high-dimensional data.

Results and discussion

The overall prediction performance of a model is measured by the area under the curve (AUC) such as ROC curve, a common evaluation metric for binary classification problems. The accuracy and F1 score are typically reported based on the threshold value of 0.5, where accuracy is the proportion of the total number of correct predictions, and the F1 score is the harmonic mean of precision and recall. The results for our data set are given in Table II. Due to the randomness in some machine learning algorithms, such as RF, and NN, the results vary even with the same data set and parameters. Therefore, we run these two algorithms ten times and report the average and standard deviation of AUC, accuracy, and F1 score based on the ten trials. The results show that RF has the best performance followed by NN and SVM. In Table II, we also report the results after discretization of the data. The AUC, accuracy, and F1 score increase dramatically for C5.0, SVM and LDA after the data are discretized, while SVM gives the highest AUC. Overall, machine learning approaches, such as RF, NN, and SVM achieve the better performance than traditional statistical methods, such as LDA and LR.

In addition to the above analysis, we also evaluated the importance of financial ratios with the LASSO model resulting in the ranking results reported in Table AII. The aim here is to see whether we can achieve better prediction performance from a model by including fewer, but relatively more important variables from each financial category. The merits of doing so include: less redundant data means less opportunity to make decisions based on noise and thus reduces overfitting; less misleading data improves modeling accuracy; less data means that algorithms train faster; and less variables provides better understanding of underlying process and making the model more interpretable.

To test this, we selected one financial ratio from each financial category, and used them to build the models. These seven selected financial ratios, according to the importance rank from LASSO, are T21500: account receivable/sales revenue, T60800 (PE ratio): market value per share/earnings per share, T40501 (return on current assets): net income/current assets, T70100 (operating cash flow ratio); operating cash flow/current liabilities, T50200 (operating leverage): gross profit/(operating profit + non-operating revenue – non-operating expenses) + finance expense, T32100 (long-term assets ratio): (total shareholders' equity + long-term liabilities)/(fixed assets + long-term investment), and T10400 (working capital ratio): (current assets-current liabilities)/current assets. The descriptive statistics of these seven financial ratios in the original data form is reported in Table AIII. A bar plot of each financial ratio against the dependent variable which is ST or non-ST company after the financial ratios were discretized to categorical variables is given in Figure A1.

Table III gives the results derived from the models based on the seven financial ratios. Our results show that the performances of these models are better than those of the models based on all 95 financial ratios as shown in Figure 1.

	Original			Discretized		
	AUC	Accuracy	F1 score	AUC	Accuracy	F1 score
RF	0.754 (0.008)	0.735 (0.008)	0.735 (0.009)	0.754 (0.007)	0.748 (0.008)	0.747 (0.006)
NN	0.739 (0.014)	0.698 (0.025)	0.676 (0.023)	0.747 (0.028)	0.721 (0.029)	0.708 (0.024)
SVM	0.709	0.661	0.632	0.765	0.742	0.742
C5.0	0.666	0.597	0.590	0.722	0.677	0.667
LDA	0.605	0.629	0.531	0.728	0.645	0.621
LOGISTIC	0.695	0.677	0.630	0.666	0.597	0.510

Table II. Prediction results for all features with original and discretized data

Moreover, the parsimonious model based on C5.0 with fewer variables enabled us to provide managers with a more concise and quantified insight because the biggest benefit of the DT model is that the output can be easily interpreted as rules. Figure 2 shows the output from C5.0 based on seven financial ratios and with original data. A set of rules is summarized below to detect financial deterioration of Chinese firms:

- Rule 1: $(T21500 > 0.712236) \rightarrow ST$
- Rule 2: $(T21500 \leqslant 0.712236, T40501 \leqslant 0.057721$, and $T60800 > 159.5) \rightarrow ST$
- Rule 3: $(T21500 \leqslant 0.712236, T40501 > 0.057721, T21500 > 0.41521$, and $T10400 \leqslant 0.032259) \rightarrow ST$
- Rule 4: $(T21500 \leqslant 0.712236, T40501 > 0.057721, T21500 > 0.41521$, and $T10400 > 0.032259$, and $T60800 \leqslant 45.9) \rightarrow ST$

Table III.
Prediction results for subset of features with original and discretized data

	Original AUC	Original Accuracy	Original F1 score	Discretized AUC	Discretized Accuracy	Discretized F1 score
RF	0.772 (0.010)	0.724 (0.014)	0.712 (0.017)	0.765 (0.008)	0.731 (0.008)	0.727 (0.009)
NN	0.756 (0.008)	0.734 (0.024)	0.721 (0.034)	0.769 (0.012)	0.729 (0.021)	0.726 (0.022)
SVM	0.759	0.726	0.721	0.769	0.742	0.750
C5.0	0.696	0.661	0.571	0.719	0.677	0.655
LDA	0.706	0.742	0.704	0.728	0.742	0.733
LOGISTIC	0.742	0.710	0.667	0.747	0.742	0.733

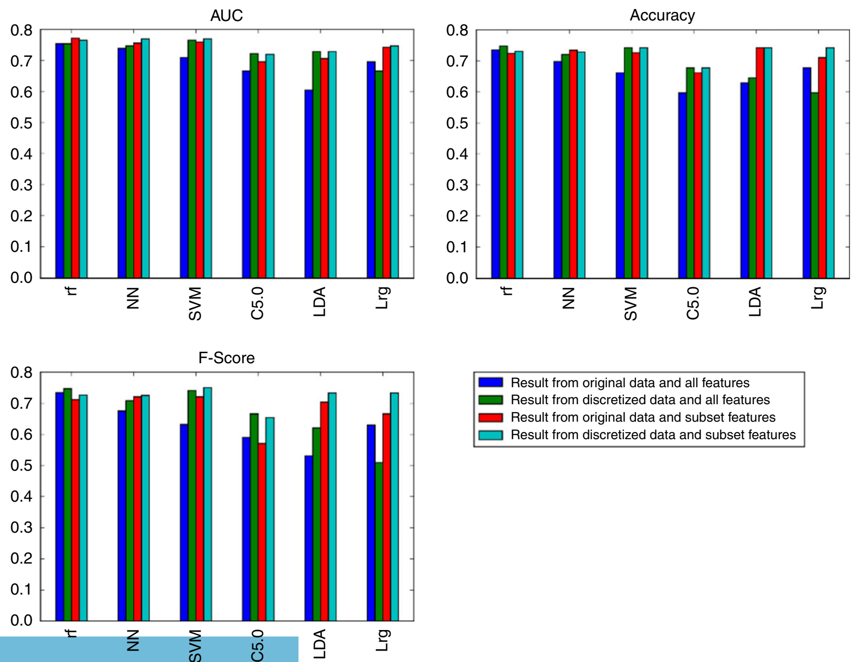

Figure 1.
Prediction results of original vs discretized and all features vs subset of features

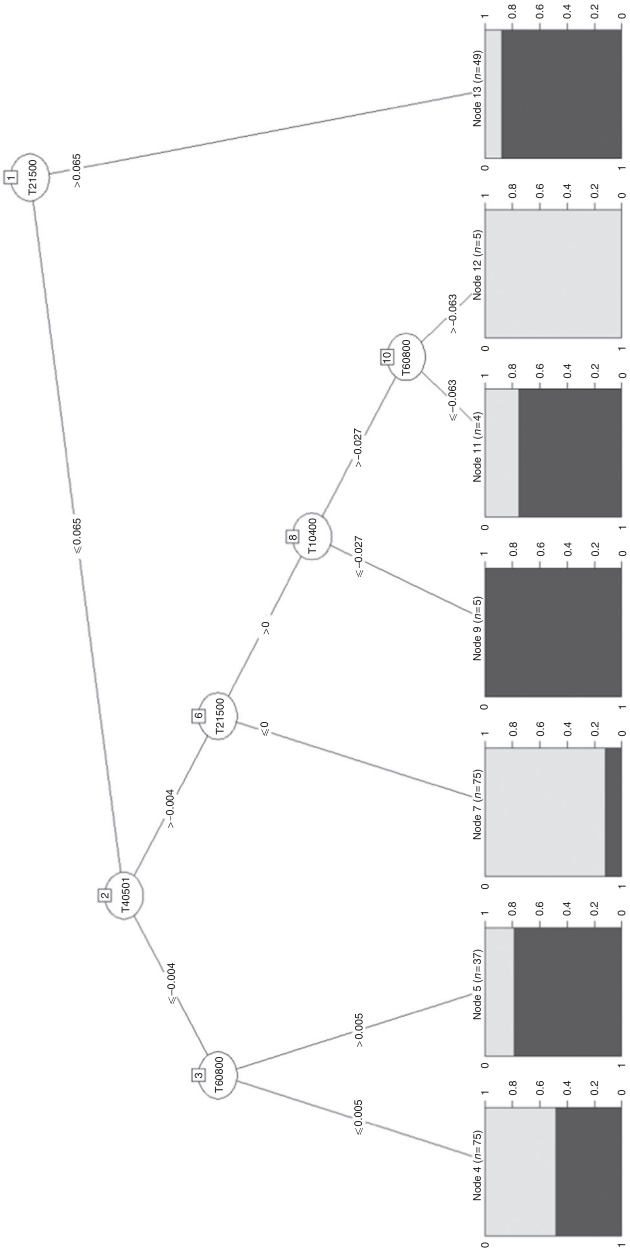

Figure 2.
Plot of decision
tree from C5.0

Among the seven financial ratios, C5.0 uses T21500 as the root node. This is due to the importance ranking from LASSO where T21500 ranks first. This financial ratio from the operating category indicates the great impact of operating capability on the prediction of financial distress for Chinese companies. A likely explanation for this is, unlike developed countries where companies have more financial sources to raise funds to mitigate distress, China is still an emerging country where many companies do not have adequate resources or easy access to commercial finance. The revenue generated from operating activities is still the most important financial source for a Chinese company. In this case, the operational efficiency to free up working capital from accounts receivable is crucial. Companies should establish a strong accounts receivable collection policy. Rule 1 states that companies have a higher chance to face financial distress when the ratio of accounts receivable over sales revenue is high. Managers should be watchful when the ratio of accounts receivable over sales revenue exceeds 0.71. Rule 2 states that when the ratio of accounts receivable over sales revenue is no more than 0.71, but the ratio of return on current assets is low, less than 0.058 and PE ratio is high, greater than 159.5, a company is likely to face financial distress in the future. Similar explanation applies to rules 3 and 4.

Conclusion

In this study, we examined the data from ST and non-ST companies from CSMAR to predict financial distress for Chinese companies. Four machine learning and two traditional statistical techniques were used as classification methods to build models based on 95 initial financial ratios, which are all continuous variables in their original form. An exciting finding from this study is that the classification accuracy increases significantly after discretizing these continuous variables. We, therefore, believe that variable discretization can be an essential pre-processing step to improve the prediction performance for classification problems which involve many continuous financial ratios.

The study also reports that machine learning approaches can achieve better performance than traditional statistical methods in classification tasks advocated by many studies (Lessmann *et al.*, 2015; West, 2000). Among the machine learning approaches for this data set, we find that RF, SVM, and NNs are the best methods for consistently predicting financial distress, and thus can be used as a tool to establish a warning mechanism so that companies can detect financial deterioration in an early stage, and make solution plans to improve their financial performance. However, due to the "blackbox" nature of these algorithms, they are unable to provide rule-based interpretation as C5.0 does. In addition, the reliability of current methods and models used in the financial industry can decrease over time due to the global economic environment (Cámská, 2015).

Finally, we apply the LASSO technique to identify the most important financial ratios from each category, and then use these ratios to build predictive models. The results show that a better prediction accuracy can be achieved by including fewer but relatively more important variables in a model. Further exploration of the top ranked ratios shows that the ratio of accounts receivable/sales revenue from operating category is a very important indicator in detecting financial distress for the companies considered here rules generated with C5.0 provide important insights for managers. They should carefully watch these ratios and promptly identify signs of financial distress for the future.

References

Alderson, M.J. and Betker, B.L. (1995), "Liquidation costs and capital structure", *Journal of Financial Economics*, Vol. 39 No. 1, pp. 45-69.

Almamy, J., Aston, J. and Ngwa, L.N. (2016), "An evaluation of Altman's Z-score using cash flow ratio to predict corporate failure amid the recent financial crisis: evidence from the UK", *Journal of Corporate Finance*, Vol. 36 No. C, pp. 278-285.

Almeida, H., Campello, M. and Weisbach, M.S. (2011), "Corporate financial and investment policies when future financing is not frictionless", *Journal of Corporate Finance*, Vol. 17 No. 3, pp. 675-693.

Altman, E.I. (1968), "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *The Journal of Finance*, Vol. 23 No. 4, pp. 589-609.

Altman, E.I., Marco, G. and Varetto, F. (1994), "Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience)", *Journal of Banking & Finance*, Vol. 18 No. 3, pp. 505-529.

Bellotti, T. and Crook, J. (2009), "Support vector machines for credit scoring and discovery of significant features", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 3302-3308.

Bhagat, S., Brickley, J.A. and Coles, J.L. (1994), "The costs of inefficient bargaining and financial distress: evidence from corporate lawsuits", *Journal of Financial Economics*, Vol. 35 No. 2, pp. 221-247.

Bolon-Canedo, V., Sanchez-Maroo, N. and Alonso-Betanzos, A. (2009), "A combination of discretization and filter methods for improving classification performance in KDD Cup 99 dataset", *International Joint Conference on Neural Networks*, *June 14-19*, pp. 359-366.

Cai, M. and Huang, Z. (2014), "Analysis of non performing loan and capital adequacy ratio among Chinese banks in the post-reform period in China", *Journal of Advanced Studies in Finance*, Vol. 5 No. 2, pp. 133-144.

Cámská, D. (2015), "Impact of the Czech changing economic environment on bankruptcy models", *International Advances in Economic Research*, Vol. 21 No. 1, pp. 117-118.

Chen, W.-S. and Du, Y.-K. (2009), "Using neural networks and data mining techniques for the financial distress prediction model", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 4075-4086.

Daily, C.M. and Dalton, D.R. (1994), "Corporate governance and the bankrupt firm: an empirical assessment", *Strategic Management Journal*, Vol. 15 No. 8, pp. 643-654.

Deangelo, H. and Deangelo, L. (1990), "Dividend policy and financial distress: an empirical investigation of troubled NYSE firms", *The Journal of Finance*, Vol. 45 No. 5, pp. 1415-1431.

Dellepiane, U., DI Marcantonio, M., Laghi, E. and Renzi, S. (2015), "Bankruptcy prediction using support vector machines and feature selection during the recent financial crisis", *International Journal of Economics and Finance*, Vol. 7 No. 8, pp. 182-195.

Demiroglu, C. and James, C. (2011), "The use of bank lines of credit in corporate liquidity management: a review of empirical evidence", *Journal of Banking & Finance*, Vol. 35 No. 4, pp. 775-782.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), "Least angle regression", *The Annals of Statistics*, Vol. 32 No. 2, pp. 407-499.

Etemadi, H., Anvary Rostamy, A.A. and Dehkordi, H.F. (2009), "A genetic programming model for bankruptcy prediction: empirical evidence from Iran", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 3199-3207.

Fedorova, E., Gilenko, E. and Dovzhenko, S. (2013), "Bankruptcy prediction for Russian companies: application of combined classifiers", *Expert Systems with Applications*, Vol. 40 No. 18, pp. 7285-7293.

Fernndez-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014), "Do we need hundreds of classifiers to solve real world classification problems?", *Journal of Machine Learning Research*, Vol. 15 No. 1, pp. 3133-3181.

Foster, B.P. and Zurada, J. (2013), "Loan defaults and hazard models for bankruptcy prediction", *Managerial Auditing Journal*, Vol. 28 No. 6, pp. 516-541.

Frydman, H., Altman, E.I. and D-L. Kao. (1985), "Introducing recursive partitioning for financial classification: the case of financial distress", *Journal of Finance*, Vol. 40 No. 1, pp. 269-291.

Gan, C., Zhang, Y., Li, Z. and Cohen, D.A. (2014), "The evolution of China's banking system: bank loan announcements 1996-2009", *Accounting & Finance*, Vol. 54 No. 1, pp. 165-188.

Geng, R., Bose, I. and Chen, X. (2015), "Prediction of financial distress: an empirical study of listed Chinese companies using data mining", *European Journal of Operational Research*, Vol. 241 No. 1, pp. 236-247.

Gestel, T.V., Baesens, B., Suykens, J.A.K., Van Den Poel, D., Baestaens, D.-E. and Willekens, M. (2006), "Bayesian kernel based classification for financial distress detection", *European Journal of Operational Research*, Vol. 172 No. 3, pp. 979-1003.

Griffiths, J.J. (2005), "The use of CDO Structuring for the disposal of non-performing loans in China", *Journal of Structured Finance*, Vol. 11 No. 3, pp. 40-50.

Guyon, I. and Elisseeff, A. (2003), "An introduction to variable and feature selection", *Journal of Machine Learning Research*, Vol. 3 Nos 7-8, pp. 1157-1182.

Härdle, W., Lee, Y.-J., Schäfer, D. and Yeh, Y.-R. (2009), "Variable selection and oversampling in the use of smooth support vector machines for predicting the default risk of companies", *Journal of Forecasting*, Vol. 28 No. 6, pp. 512-534.

Hilston Keener, M. (2013), "Predicting the financial failure of retail companies in the United States", *Journal of Business & Economics Research* (*Online*), Vol. 11 No. 8, pp. 373-380.

Hoshi, T., Kashyap, A. and Scharfstein, D. (1990), "The role of banks in reducing the costs of financial distress in Japan", *Journal of Financial Economics*, Vol. 27 No. 1, pp. 67-88.

Jain, B.A. and Nag, B.N. (1997), "Performance evaluation of neural network decision models", *Journal of Management Information Systems*, Vol. 14 No. 2, pp. 201-216.

John, T.A. (1993), "Accounting measures of corporate liquidity, leverage, and costs of financial distress", *Financial Management*, Vol. 22 No. 3, pp. 91-100.

Karami, G., Hosseini, S.M.S., Attaran, N. and Hosseini, S.M.S. (2012), "Bankruptcy prediction using memetic algorithm with fuzzy approach: empirical evidence from Iran", *International Journal of Economics and Finance*, Vol. 4 No. 5, pp. 116-123.

Kaul, A. (2014), "Lasso with long memory regression errors", *Journal of Statistical Planning and Inference*, Vol. 153, pp. 11-26.

Keasey, K. and Watson, R. (1991), "Financial distress prediction models: a review of their usefulness", *British Journal of Management*, Vol. 2 No. 2, pp. 89-102.

Korol, T. (2013), "Early warning models against bankruptcy risk for central European and Latin American enterprises", *Economic Modelling*, Vol. 31, pp. 22-30.

Kumar, P.R. and Ravi, V. (2007), "Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review", *European Journal of Operational Research*, Vol. 180 No. 1, pp. 1-28.

Kwak, W., Shi, Y. and Kou, G. (2012), "Bankruptcy prediction for Korean firms after the 1997 financial crisis: using a multiple criteria linear programming data mining approach", *Review of Quantitative Finance and Accounting*, Vol. 38 No. 4, pp. 441-453.

Lacher, R.C., Coats, P.K., Sharma, S.C. and Fant, L.F. (1995), "A neural network for classifying the financial health of a firm", *European Journal of Operational Research*, Vol. 85 No. 1, pp. 53-65.

Lawrence, E.C. and Bear, R.M. (1986), "Corporate bankruptcy prediction and the impact of leases", *Journal of Business Finance & Accounting*, Vol. 13 No. 4, pp. 571-585.

Lee, K.C., Han, I. and Kwon, Y. (1996), "Hybrid neural network models for bankruptcy predictions", *Decision Support Systems*, Vol. 18 No. 1, pp. 63-72.

Lee, T.S. and Chen, I.F. (2005), "A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines", *Expert Systems with Applications*, Vol. 28 No. 4, pp. 743-752.

Leshno, M. and Spector, Y. (1996), "Neural network prediction analysis: the bankruptcy case", *Neurocomputing*, Vol. 10 No. 2, pp. 125-147.

Lessmann, S., Baesens, B., Seow, H.-V. and Thomas, L.C. (2015), "Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research", *European Journal of Operational Research*, Vol. 247 No. 1, pp. 124-136.

Li, Z., Crook, J. and Andreeva, G. (2014), "Chinese companies distress prediction: an application of data envelopment analysis", *Journal of the Operational Research Society*, Vol. 65 No. 3, pp. 466-479.

Lian, R. (2016), "Chinese Banks' New Bad Loans more than Doubled in 2015", Reuters, Shanghai, available at: www.reuters.com/article/us-china-banking-npl-idUSKCN0UQ04D20160112 (accessed January 12, 2016).

Liang, D., Lu, C.-C., Tsai, C.-F. and Shih, G.-A. (2016), "Financial ratios and corporate governance indicators in bankruptcy prediction: a comprehensive study", *European Journal of Operational Research*, Vol. 252 No. 2, pp. 561-572.

Lu, D., Thangavelu, S. and Hu, Q. (2005), "Biased lending and non-performing loans in China's banking sector", *Journal of Development Studies*, Vol. 41 No. 6, pp. 1071-1091.

Lustgarten, J.L., Gopalakrishnan, V., Grover, H. and Visweswaran, S. (2008), "Improving classification performance with discretization on biomedical datasets", *AMIA Annual Symposium Proceedings*, pp. 445-449.

Luther, R.K. (1998), "An artificial neural network approach to predicting the outcome of chapter 11 bankruptcy", *The Journal of Business and Economic Studies*, Vol. 4 No. 1, pp. 57-73.

Mckee, T.E. and Lensberg, T. (2002), "Genetic programming and rough sets: a hybrid approach to bankruptcy classification", *European Journal of Operational Research*, Vol. 138 No. 2, pp. 436-451.

Miller, A.J. (1984), "Selection of subsets of regression variables", *Journal of the Royal Statistical Society. Series A (General)*, Vol. 147 No. 3, pp. 389-425.

Min, J.H. and Jeong, C. (2009), "A binary classification method for bankruptcy prediction", *Expert Systems with Applications*, Vol. 36 No. 3, pp. 5256-5263.

Olson, D.L., Delen, D. and Meng, Y. (2012), "Comparative analysis of data mining methods for bankruptcy prediction", *Decision Support Systems*, Vol. 52 No. 2, pp. 464-473.

Opler, T.C. and Titman, S. (1994), "Financial distress and corporate performance", *The Journal of Finance*, Vol. 49 No. 3, pp. 1015-1040.

Osborn, T., Jong, V., So, C. and Dilley, J. (2015), *China's Non-Performing Loans are Rising Fast*, PWC, Hong Kong.

Pendharkar, P.C. (2005), "A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem", *Computers & Operations Research*, Vol. 32 No. 10, pp. 2561-2582.

Platt, H.D. and Platt, M.B. (2002), "Predicting corporate financial distress: reflections on choice-based sample bias", *Journal of Economics and Finance*, Vol. 26 No. 2, pp. 184-199.

Porzecanski, K. (2016), "Bass says China bank losses may top 400% of subprime crisis Bloomberg business, available at: www.bloomberg.com/news/articles/2016-02-10/bass-says-china-s-banking-losses-may-top-400-of-subprime-crisis (accessed November 10, 2016).

Potena, P. (2013), "Optimization of adaptation plans for a service-oriented architecture with cost, reliability, availability and performance tradeoff", *Journal of Systems and Software*, Vol. 86 No. 3, pp. 624-648.

Ryu, Y.U. and Yue, W.T. (2005), "Firm bankruptcy prediction: experimental comparison of isotonic separation and other classification approaches", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 35 No. 5, pp. 727-737.

Senteney, D.L., Bazaz, M.S. and Ahmadpour, A. (2006), "Tests of the incremental explanatory power of auditor qualified opinion and audit firm changes in predicting impending bankruptcy", *International Journal of Accounting, Auditing and Performance Evaluation*, Vol. 3 No. 4, pp. 434-451.

Shin, K.-S., Lee, T.S. and Kim, H.-J. (2005), "An application of support vector machines in bankruptcy prediction model", *Expert Systems with Applications*, Vol. 28 No. 1, pp. 127-135.

Sinha, A.P. and Zhao, H. (2008), "Incorporating domain knowledge into data mining classifiers: an application in indirect lending", *Decision Support Systems*, Vol. 46 No. 1, pp. 287-299.

Sudarsanam, S. and Lai, J. (2001), "Corporate financial distress and turnaround strategies: an empirical analysis", *British Journal of Management*, Vol. 12 No. 3, pp. 183-199.

Sun, J., Li, H., Huang, Q.-H. and 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*, Vol. 57, pp. 41-56.

Suzuki, Y., Miah, D. and Jinyi, Y. (2008), "China's non-performing bank loan crisis: the role of economic rents", *Asian-Pacific Economic Literature*, Vol. 22 No. 1, pp. 57-70.

Theodossiou, P., Kahya, E., Saidi, R. and Philippatos, G. (1996), "Financial distress and corporate acquisitions: further empirical evidence", *Journal of Business Finance & Accounting*, Vol. 23 Nos 5-6, pp. 699-719.

Tian, S., Yu, Y. and Guo, H. (2015), "Variable selection and corporate bankruptcy forecasts", *Journal of Banking & Finance*, Vol. 52, pp. 89-100.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58 No. 1, pp. 267-288.

Tillander, A. (2012), "Effect of data discretization on the classification accuracy in a high-dimensional framework", *International Journal of Intelligent Systems*, Vol. 27 No. 4, pp. 355-374.

Tsai, C.-F. (2008), "Financial decision support using neural networks and support vector machines", *Expert Systems*, Vol. 25 No. 4, pp. 380-393.

Tsai, C.-F. and Hsu, Y.-F. (2013), "A meta-learning framework for bankruptcy prediction", *Journal of Forecasting*, Vol. 32 No. 2, pp. 167-179.

West, D. (2000), "Neural network credit scoring models", *Computers & Operations Research*, Vol. 27 Nos 11-12, pp. 1131-1152.

Whiting, D.G., Hansen, J.V., Mcdonald, J.B., Albrecht, C. and Albrecht, W.S. (2012), "Machine learning methods for detecting patterns of management fraud", *Computational Intelligence*, Vol. 28 No. 4, pp. 505-527.

Wuerges, A.F.E. and Borba, J.A. (2010), "Redes neurais, lógica nebulosa e algoritmos genéticos: aplicações e possibilidades em finanças e contabilidade/neural networks, fuzzy logic and genetic algorithms: applications and possibilities in finance and accounting", *Journal of Information Systems and Technology Management*, Vol. 7 No. 1, pp. 163-182.

Yang, Z., You, W. and Ji, G. (2011), "Using partial least squares and support vector machines for bankruptcy prediction", *Expert Systems with Applications*, Vol. 38 No. 7, pp. 8336-8342.

Yang, Z.R., Platt, M.B. and Platt, H.D. (1999), "Probabilistic neural networks in bankruptcy prediction", *Journal of Business Research*, Vol. 44 No. 2, pp. 67-74.

Youn, H. and Gu, Z. (2010), "Predict US restaurant firm failures: the artificial neural network model versus logistic regression model", *Tourism and Hospitality Research*, Vol. 10 No. 3, pp. 171-187.

Zavgren, C.V., Dugan, M.T. and Reeve, J.M. (1988), "The association between probabilities of bankruptcy and market responses – a test of market anticipation", *Journal of Business Finance & Accounting*, Vol. 15 No. 1, pp. 27-45.

Zha, Y., Liang, N., Wu, M. and Bian, Y. (2016), "Efficiency evaluation of banks in China: a dynamic two-stage slacks-based measure approach", *Omega*, Vol. 60, pp. 60-72.

Zhang, D., Cai, J., Dickinson, D.G. and Kutan, A.M. (2016), "Non-performing loans, moral hazard and regulation of the Chinese commercial banking system", *Journal of Banking & Finance*, Vol. 63, pp. 48-60.

Zhang, G., Hu, M.Y., Patuwo, B.E. and Indro, D.C. (1999), "Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis", *European Journal of Operational Research*, Vol. 116 No. 1, pp. 16-32.

Zhu, N., Wang, B. and Wu, Y. (2015), "Productivity, efficiency, and non-performing loans in the Chinese banking industry", *Social Science Journal*, Vol. 52 No. 4, pp. 468-480.

Appendix 1

Table AI. List of financial indicators

Code	Financial indicators	Category
T10100	Current assets/current liabilities	Liquidity (4/95)
T10300	Cash and cash equivalents/current liabilities	
T10400	(Current assets-current liabilities)/current assets	
T10600	Working capital/net total assets	
T20101	Sales revenue/account receivable at the end of year	Operating (14/95)
T20201	Accounts receivable turnover days	
T20301	Operating costs/account payable	
T20401	Sales revenue/working capital	
T20601	Operating costs/current assets at the end year	
T20701	Operating costs/fixed assets at the end year	
T20801	Operating costs/long-term assets at the end year	
T20901	Operating costs/total assets at the end year	
T21001	Sales revenue/shareholder's equity	
T21100	Total assets/sales revenue	
T21500	Account receivable/sales revenue	
T21600	Inventory/sales revenue	
T21700	Current assets/operating costs	
T21800	Fixed assets/operating costs	
T30100	Total liabilities/working capital	Solvency (20/95)
T30200	Total shareholders' equity/total assets	
T30300	Current assets/total assets	
T30400	Fixed assets/total assets	
T30500	Total shareholders' equity/fixed assets	
T30600	Current liabilities/total liabilities	
T30700	Long-term liabilities/total liabilities	
T30800	Total shareholders' equity/total liabilities	
T30900	Total liabilities/total tangible assets	
T31000	Total liabilities/market price	
T31101	(Net income + income tax + financial expenses)/financial expenses	
T31300	Total liabilities/shareholder's equity	
T31400	Total assets/shareholder's equity	
T31500	Non-current liabilities/(non-current liabilities + shareholder's equity)	
T31800	(Total assets-current assets)/total assets	
T31900	Tangible assets/total assets	
T32100	(Total shareholders' equity + long-term liabilities)/(fixed assets + long-term investment)	
T32200	Working capital/(short-term debt + long-term debt)	
T32300	Long-term debt/total assets	
T40100	(Sales revenue-operating costs)/sales revenue	Profitability (22/95)
T40200	Net income/sales revenue	
T40301	(Gross profit + financial expenses)/total assets	
T40401	Net income/total assets (ROA)	
T40501	Net income/current assets	
T40601	Net income/fixed assets	
T40801	Net income/shareholder's equity (ROE)	
T40900	Operating profit/sales revenue	
T40901	Income tax/gross profit	
T41200	(Net income + financial expenses)/(total assets-current liabilities + notes payable + short-term debt + long-term debt due in 1 year)	
T41300	Sales tax/sales revenue	

(continued)

Table AI.

Code	Financial indicators	Category
T41400	Operating costs/sales revenue	
T41500	(Sales expenses + administration expenses + financial expenses)/ sales revenue	
T41600	Gross profit/(operating costs + sales expenses + administration expenses + financial expenses	
T41700	(Gross profit + financial expenses)/(average long-term debt + average shareholder's equity)	
T41800	Net income/gross income	
T41900	Gross income/EBIT	
T42000	EBIT/sales revenue	
T42100	Sales expenses/sales revenue	
T42200	Administration expenses/sales revenue	
T42300	Finanical expenses/sales revenue	
T42500	EBIT/total assets	
T50100	(Gross profit + financial expenses)/gross profit	Leverage (2/95)
T50200	(Sales revenue-operating costs)/(net income + financial expense)	
T60100	Sales revenue/total shares	Shareholders' profitability (19/95)
T60200	Net income/total shares	
T60300	Total shareholders' equity/common Shares Issued	
T60400	Market value per share/net assets per share	
T60500	Surplus reserves/total shares	
T60600	Capital reserves/total shares	
T60700	Undistributed profit/total shares	
T60800	Market value per share/earnings per share	
T61102	Dividend per share + market value of stock at beginning of the year – market value of stock at the end of the year)/market value per share	
T61300	Share price/cash flow per share	
T61400	Share price/revenue per share	
T61601	Total market value (A)/total assets at the end of year	
T61701	Total assets at the end of year/total market value (A)	
T61800	(Surplus reserves + undistributed profit)/total assets	
T62000	Shareholder's equity/invested capital	
T62100	EBIT/total shares	
T62200	Retained earnings/total shares	
T62300	Free cash flow for the firm/number of share of stock	
T62400	Free cash flow of equity/number of share of stock	
T70100	Operating cash flow/current liabilities	Cash flow (14/95)
T70200	Operating cash flow/operation revenue	
T70300	Operating cash flow/total shares	
T70400	Investment activities net cash flow/total shares	
T70500	Financing activities net cash flow/total shares	
T70600	Net increase in cash and cash equivalents/total shares	
T71800	Net cash flow from operating/net income	
T71900	Net cash flow from operating/gross profit	
T72000	Net cash flow from operating/financial expenses	
T72100	Operating cash flow/total liabilities	
T72200	Net cash flow from operating/(long-term debt due in 1 year + notes payable)	
T72500	(Net cash flow from operating – cash dividends – interest expense)/(fixed assets + investment + working capital)	
T72700	Operating cash flow/total assets	
T73000	Cash received/operation revenue	

Appendix 2

1. T21500	2. T60800	3. T40501	4. T41200	5. T32100	6. T70100
7. T71800	8. T41300	9. T41900	10. T70600	11. T21001	12. T40801
13. T42300	14. T60500	15. T72200	16. T20401	17. T50200	18. T20201
19. T30500	20. T61102	21. T41400	22. T21100	23. T20101	24. T30800
25. T30900	26. T61701	27. T72100	28. T42200	29. T31101	30. T30400
31. T73000	32. T40100	33. T30600	34. T30700	35. T60600	36. T61300
37. T10400	38. T42000	39. T61400	40. T42500	41. T20601	42. T32300
43. T31900	44. T70200	45. T10600	46. T10300	47. T50100	48. T60700
49. T72700	50. T10100	51. T40901	52. T41700	53. T42100	54. T62400
55. T20301	56. T21700	57. T61800	58. T31000	59. T40301	60. T61601
61. T72500	62. T20801	63. T72000	64. T31500	65. T32200	66. T41600
67. T60100	68. T20901	69. T60400	70. T62000	71. T62300	72. T31300
73. T70400	74. T30300	75. T31800	76. T50300	77. T62100	78. T21800
79. T40900	80. T70300	81. T40401	82. T40200	83. T60200	84. T70500
85. T21600	86. T30100	87. T30200	88. T31400	89. T60300	90. T40601
91. T41800	92. T71900	93. T62200	94. T20701	95. T41500	

Table AII. Importance ranking from LASSO

Appendix 3

	T21500	T60800	T40501	T70100	T50200	T32100	T10400
Mean	0.414	150.648	0.068	0.133	−199.044	2.048	0.112
Std	0.532	229.623	0.144	0.251	3559.032	2.047	0.477
Min.	0.002	7.062	−0.990	−0.719	−62862.552	0.285	−2.081
25%	0.109	32.619	0.016	0.015	1.502	1.185	−0.084
50%	0.223	65.567	0.049	0.087	2.059	1.638	0.226
75%	0.532	162.230	0.092	0.202	2.956	2.349	0.432
Max.	4.587	1670.000	1.431	1.384	33.194	27.479	0.872

Table AIII. Descriptive statistics of seven selected financial ratios based on original data

Appendix 4

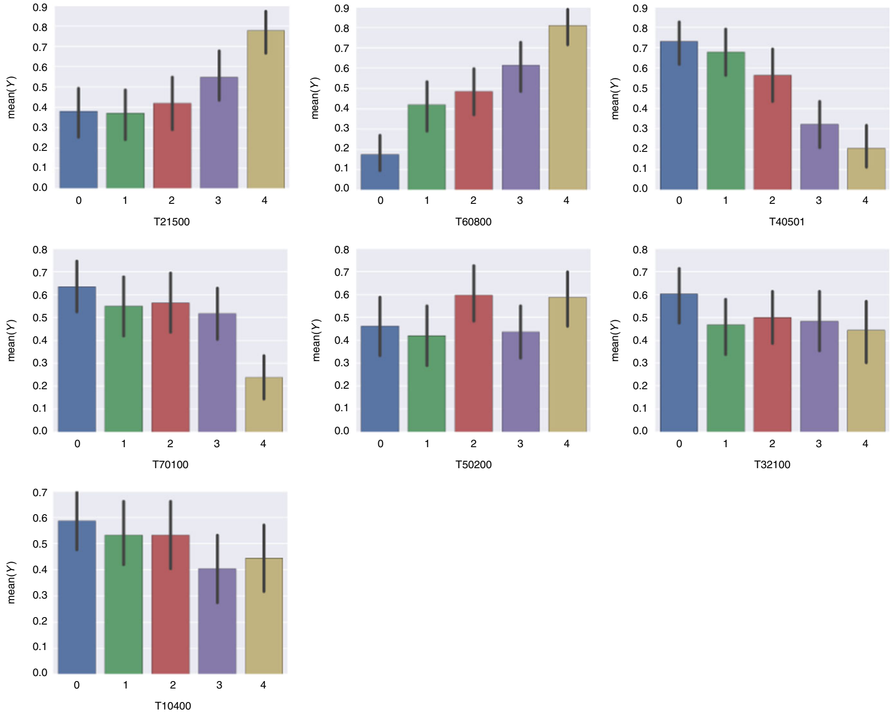

Figure A1.
Bar plot of each financial ratios against the dependent variable

About the authors

Jun Huang is currently an Assistant Professor of Management in Department of Management and Marketing, College of Business, Angelo State University. He received a PhD Degree in International Business with Quantitative Management concentration from Texas A&M International University. His research interests include data mining in business analytics, feature selection for high-dimensional data and its application in prediction and classification problems, optimization problems in management, and international business/management. He has publication in such outlets as *Communication in Statistics, Journal of Modelling in Management, International Journal of Management and Decision Making and Advanced Materials Research*, and proceeding papers and conference presentations with INFORMS, IEEE, etc.

Haibo Wang is a Radcliff Killam Distinguished Professor in Operation Research and Decision Sciences at AR Sanchez Jr School of Business, Texas A&M International University, he received a PhD Degree in Production Operations Management in 2004 from The University of Mississippi. He is a Guest Editor and Editorial Board Member of several international journals. He has publications in such outlets as *European Journal of Operational Research, Journal of Intelligent and Robotic Systems, Computers and Operation Research, IEEE transactions on Control System Technology, IEEE transactions on Automation Science and Engineering, Journal of Operational Research Society, Computers and Industrial Engineering, Journal of Applied Mathematical Modeling, International Journal of Flexible Manufacturing Systems, International Journal of Production Research, Journal of Human and Ecological Risk Assessment, Journal of Heuristics, Communications in Statistics, International Journal of Information Technology and Decision Making, Journal of Combinatorial Optimization, Journal of Optimization Letters*, etc. Haibo Wang is the corresponding author and can be contacted at: hwang@tamiu.edu

Dr Gary Kochenberger is a Professor of Business Analytics at the University of Colorado, Denver where he is a Co-director of the Business Analytics graduate program. He is known for his research on applied optimization addressing important issues in combinatorial optimization, nonlinear programming, resource allocation, pattern classification, data mining, and related areas. He has co-authored three books and more than 70 refereed articles. Dr Kochenberger has given invited presentations at national and international meetings in the USA, Canada, China, Japan, and several venues in Europe. He has also served as a principal investigator and supporting investigator on numerous grants. In addition to his academic work, Dr Kochenberger serves as a Senior Consultant for OptTek Systems, Inc., engaged in research into practical applications of optimization and simulation applied to portfolio analysis, workforce planning, and a variety of other applications of operations research.