Goo et al. SpringerPlus (2016) 5:539
DOI 10.1186/s40064-016-2186-5

SpringerPlus

RESEARCH Open Access

Improving the prediction of going concern of Taiwanese listed companies using a hybrid of LASSO with data mining techniques

Yeung-Ja James Goo[1], Der-Jang Chi[2*] and Zong-De Shen[1]

*Correspondence:
qq1011108@gmail.com
[2] Department of Accounting, Chinese Culture University, No. 55, Hwa-Kang Road, Yang-Ming-Shan, Taipei City 11114, Taiwan
Full list of author information is available at the end of the article

Abstract

The purpose of this study is to establish rigorous and reliable going concern doubt (GCD) prediction models. This study first uses the least absolute shrinkage and selection operator (LASSO) to select variables and then applies data mining techniques to establish prediction models, such as neural network (NN), classification and regression tree (CART), and support vector machine (SVM). The samples of this study include 48 GCD listed companies and 124 NGCD (non-GCD) listed companies from 2002 to 2013 in the TEJ database. We conduct fivefold cross validation in order to identify the prediction accuracy. According to the empirical results, the prediction accuracy of the LASSO–NN model is 88.96 % (Type I error rate is 12.22 %; Type II error rate is 7.50 %), the prediction accuracy of the LASSO–CART model is 88.75 % (Type I error rate is 13.61 %; Type II error rate is 14.17 %), and the prediction accuracy of the LASSO–SVM model is 89.79 % (Type I error rate is 10.00 %; Type II error rate is 15.83 %).

Keywords: Going concern prediction, Least absolute shrinkage and selection operator (LASSO), Data mining, Neural network (NN), Classification and regression tree (CART), Support vector machine (SVM)

Background

Business bankruptcy has caused a huge loss of wealth on the part of investors. Hence, building a valid going concern problem forecast model for an enterprise has become an important goal for both academics and financial practitioners. The high association between going concern doubts (GCD) and business bankruptcy has been verified by past studies (Behn et al. 2001; Geiger and Rama 2003; Koh and Low 2004; Martens et al. 2008; Mokhatab et al. 2011; Yeh et al. 2014). Moreover, the Statement of Auditing Standard (SAS) demands that when an auditor suspects the auditee's capability of going concern, the auditor should conduct the necessary and reasonable auditing processes required to examine the auditee's related financial information. If an auditor makes a misjudgment during the auditing process and issues an incorrect audit report, then this has important consequences (e.g. business crisis or investment losses). As a result, the question of how to help auditors notice signs of going concern is an important one.

GCD and bankruptcy forecasts have over the past decade become recognizable with classification problems. Generally, the classification problem carries out a computation

in light of the numerical value of some given classification data in order to acquire the relevant classification rule for every classification, bringing unknown classification data into the rule in order to acquire the final classification result. Many going concern prediction (GCP) studies have applied neural network (NN) to build classification models and to acquire results for going concern (GC) issues (Chen and Church 1992; Cornier et al. 1995; Mutchler et al. 1997; Foster et al. 1998; Carcello and Neal 2000; Gaganis et al. 2007; Chen and Lee 2015).

In terms of statistical tools used to handle mega data analysis, machine learning has risen sharply in recent years. It identifies unknown information from complex data and aims to recognize data in order to draw an inference from the structured model, which can act as a reference amount when making decisions for different purposes that are often related to GC issues (Lenard et al. 1995; Anandarajan and Anandarajan 1999; Brabazon and Keenan 2004; Gaganis et al. 2007, Martens et al. 2008; Kirkos et al. 2007a, b; Mokhatab et al. 2011; Salehi and Fard 2013; Yeh et al. 2014; Chen and Lee 2015). The classification method is used most often in these studies, and its results are able to serve as the basis for both decisions and forecasts. However, whether any of the machine learning algorithms in GCP studies is more suitable to this task than another method remains disputed.

Aside from accuracy of the prediction models, the occurrence of Type I error and Type II error cannot be ignored (O'Leary 1998; Kirkos et al. 2007a, b; Tasi and Huang 2010; Chen et al. 2015). A Type II error may especially cause damages and high costs. If an auditor issues a wrong audit report due to his/her misjudgment, then it affects not only the enterprise and stakeholders, but also many investors. Moreover, the CPA may be sued. The costs for Type II errors are rather severe in the U.S. Examples include the Enron scandal in 2001 (Benston and Hartgraves 2002) and WorldCom fraud in 2003. Taiwan has had its own financial fraud cases for Procomp Informatics and Infodisc in 2004 and Summit Computer in 2006.

The purpose of this study is to develop a satisfactory model for forecasting the GCD of firms and to forecast an omen for such GCD and to reduce damage to both investors and auditors. This study applies support vector machine (SVM), as well as the classification and regression trees (CARTs) in the machine learning method, as its basis and matches LASSO in order to separately establish a classification model and draw up a comparison.

Literature review

Going concern concept and reports

Before investors invest in a company, they should understand the viability of the company. This kind of viability relates to the ability of management to properly manage the company's overall resources in order to survive. In uncertain situations, investors expect auditors to provide early warnings of business failure and risks of bankruptcy (Chen and Church 1996).

Pursuant to the provision of SAS No. 59, an auditor's consideration of an entity's ability to continue as GC requires an explicit evaluation of the auditee's continued viability during the audit process. As a result, the GCD report is used as a warning sign when an auditor suspects an auditee's weakness in terms of GCD (Lenard et al. 1995).

Criteria for issuing an audit report by CPA for going concern

Taiwan's auditing standards bulletin No. 16 stipulates that the compilation of financial statements is often based on an assumption of going concern. It further requires that auditors shall comply with the stipulations as specified in the bulletin when they evaluate reasonable assumptions of going concern. CPAs are able to issue unqualified opinion audit reports if they eliminate their doubt about the ability of going concern after evaluating the rationality of the assumption of going concern. If CPAs consider the auditee's future measures are reasonable and necessary to be disclosed in the financial report, then a qualified opinion audit report or an adverse opinion audit report is needed. If the CPA cannot eliminate doubts about the auditee's ability of going concern, but the auditee's financial statements have been disclosed, then the CPA shall issue an unqualified-modified opinion audit report. If the auditee's financial statements have not been properly disclosed, then the CPA shall issue a qualified opinion audit report or an adverse opinion audit report depending on the significance. If a CPA has confirmed that the assumption of going concern for the compilation of financial statements is not consistent with the actual situation and would have serious consequences, then the CPA shall issue an adverse audit opinion report. If the CPA cannot eliminate doubt, or the assumption is not consistent with the actual situation, then explanatory notes should be included in the audit report, and these notes should form the audit report (Auditing Standards Board of the Republic of China Accounting Research Development Foundation, Auditing standard bulletin and auditing practice, 2013).

Traditional classification studies

The GCP model carries out a computation that mainly depends on the numerical values of train subset data of financial and non-financial indicators in order to acquire the relevant classification rule for every classification and brings data subsets into the rule in order to acquire the final classification result.

Based on the difficulty of the GCD assessment, many authors apply LR in order to make a GCP classification in relation to the GC issue (Chen and Church 1992; Cornier et al. 1995; Mutchler et al. 1997; Foster et al. 1998; Carcello and Neal 2000; Gaganis et al. 2007). However, the traditional classification method suffers from the limitation of having to be in accordance with specific assumptions in the data.

Machine learning classification methods

The machine learning approach has often been adopted in the literature. Many studies have attempted to apply the machine learning approach as a base to build a classification model. These studies point out that adopting this method leads to outstanding prediction accuracy. Several studies applying a machine learning approach (e.g. SVM, DT, NN, etc.) to GCD, indicating that these approaches are able to forecast the GC status of businesses and provide useful financial data for the GC issue (Brabazon and Keenan 2004; Koh and Low 2004; Martens et al. 2008; Mokhatab et al. 2011; Salehi and Fard 2013; Yeh et al. 2014).

On a similar classification issue, Tasi and Wu (2008) apply NN in relation to bankruptcy predictions and credit scores. Chen et al. (2014) employ DT, SVM, and LR in the Fraudulent Financial Statements forecast in order to acquire excellent classification

results. Based on these studies, this study utilizes the aforementioned LR, SVM, NN, and DT approaches as the basis upon which to build a classification model.

Methods

The purpose of this study is to establish a two-stage going concern doubt prediction model that integrates financial and non-financial indicators. The process of this study creates a least absolute shrinkage and selection operator (LASSO) to obtain the results for important indicators of GCD after screening. For forecast modeling, the classification approach includes the following machine learning techniques: NN, DT, and SVM. Finally, this study draws a comparison and conducts an analysis in order to obtain better GC prediction results.

Least absolute shrinkage and selection operator (LASSO)

Stepwise regression has been applied in related work in the past, but there are significant problems with stepwise methods, which have been admirably summarized by Harrell (2001). These problems are as follows: (1) R^2 values are biased. (2) The F test statistics do not have the claimed distribution. (3) The standard errors of the parameter estimates are too small. (4) Consequently, the confidence intervals around the parameter estimates are too narrow. (5) The parameter estimates are highly biased in absolute value. (6) Collinearity problems are exacerbated.

This study applies LASSO as a feature selection method, which was first proposed by Tibshirani (1996). This algorithm minimizes the residual sum of squares subject to the sum of the absolute values of the coefficient being less than a constant.

$$\hat{\beta}^{L}_{\sim} = \arg\min\left\{\sum_{i=1}^{N}\left(y_1 - \alpha - \sum_{j}\beta_j \chi_{ij}\right)^2\right\} \tag{1}$$

subject to

$$\sum_{j=1}^{p}\left|\hat{\beta}_j^{L}\right| \le (Constant) \tag{2}$$

If $t > \sum_{j=1}^{p}\left|\hat{\beta}_j^{0}\right|$, then the LASSO algorithm yields the same estimate as the OLS estimate.

However, if $0 < t < \sum_{j=1}^{p}\left|\hat{\beta}_j^{0}\right|$, then the problem is equivalent to:

$$\hat{\beta}^{L}_{\sim} = \arg\min\left[\sum_{i=1}^{N}\left(y_1 - \alpha - \sum_{j}\beta_j \chi_{ij}\right)^2 + \lambda\sum_{j}\left|\beta_j\right|\right] \tag{3}$$

where, $\lambda > 0$. We shall show later that the relation between λ and the LASSO parameter t is one-to-one.

Due to the nature of the constraint, LASSO tends to produce some coefficients that are exactly zero. Compared to OLS, whose predicted coefficient $\beta^{0}_{\sim}$ is an unbiased estimator

of β, both ridge regression and LASSO sacrifice a little bias in order to reduce the variance of the predicted values and improve the overall prediction accuracy. In this past decade, LASSO has been widely applied in many different ways and variants (Tibshirani et al. 2005; Colombani et al. 2013; Yamada et al. 2014; Toiviainen et al. 2014; Connor et al. 2015).

Neural networks (NN)

Neural networks refer to information processing systems that simulate bio-neural networks. They use a large number of connected artificial neurons in order to simulate the capacity of neural networks (Anandarajan and Anandarajan 1999; Tasi and Wu 2008; Korol 2013; Chen et al. 2015). Since NN is equipped with the functions of high-speed calculation and information de-noises, it is capable of solving many sophisticated classification and forecasting issues. The most common NN model has three layers: input layer, hidden layer, and output layer. The input layer is used to receive variables. The hidden layer is constituted by neutrons, and its major purpose is to increase the complexity of neural networks, so that they can simulate complicated linear relations. The output layer generates post-processing prediction results. The three layers of the NN model are illustrated in Fig. 1.

The MLP network is a function of one or more predictors that minimizes the prediction error of one or more targets. Predictors and targets can be a mix of categorical and continuous fields. The general architecture for MLP networks can be described as:

$$\text{Input layer: } J_0 = P \text{ units, } a_{0:1,\ldots,}a_{0:j0};\ \text{with } a_{0:j0} = x_j \tag{4}$$

$$\text{ith hidden layer: } J_i \text{ units, } a_{i:1,\ldots,}a_{i:J_i};\ \text{with } a_{i:k} = \gamma_i(c_{i:k})$$
$$\text{and } c_{i:k} = \sum_{j=0}^{J_1} w_{i:j,k}a_{i-1:j},\ \text{and } a_{i-1:0} = 1 \tag{5}$$

$$\text{Output layer: } J_I = R \text{ units, } a_{I:1,\ldots,}a_{I:J_i};\ \text{with } a_{I:k} = \gamma_I(c_{I:k})$$
$$\text{and } c^m_{I:k} = \sum_{j=0}^{J_1} w_{I:j,k}a_{i-1:j},\ \text{and } a_{i-1:0} = 1 \tag{6}$$

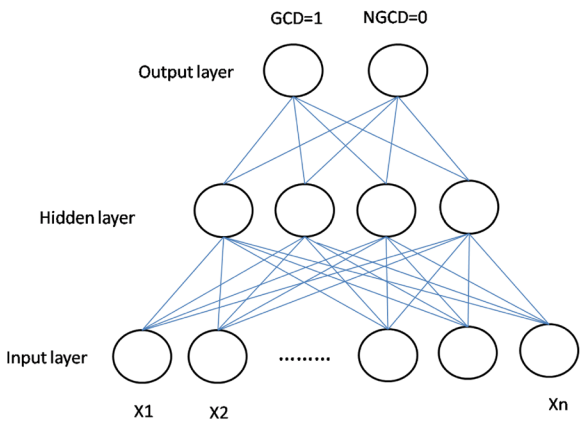

Fig. 1 Neural network model

The training finally proceeds through at least one complete pass of the data. The search should then be stopped according to the stopping criteria.

Where, $X(m) = x_{1,\ldots,}^{(m)} x_p^{(m)}$ is the input vector; pattern m, $m = 1, \ldots$ M; $Y(m) = y_{1,\ldots,}^{(m)} y_R^{(m)}$ is the target vector; pattern m; I is the number of layers, discounting the input layer; J_i is the number of units in layer i; $J_0 = P, J_i = R$, discounting the bias unit; Γ^c and Γ are a set of categorical outputs and continuous outputs; Γ_h is a set of sub-vectors of $Y^{(m)}$ containing 1-of c coded hth categorical field; and $w_{i:j,k}$ is a weight leading from layer $i-1$, unit j to layer i, unit k. No weights connect $a_{i-1:j}^m$ and the bias $a_{i:0}^m$—that is, there is no $w_{i:j,0}$ for any j. Finally, $c_{i:k}^m$ is $\sum_{j=0}^{J_i-1} w_{i:j,k} a_{i-1:j}^m$, $i = 1, \ldots, I$ and $\gamma_i(c)$ is an activation function for layer i.

Support vector machine (SVM)

Support vector machine (SVM) was developed by Boser et al. (1992) to provide better solutions than other traditional classifiers, such as neural networks. SVM is a type of maximal margin classifier, in which the classification problem can be represented as an optimization process, which finds the maximum-margin hyper-plane from a given training dataset D as described by:

$$D = \left\{ (x_i, y_i) \middle| x_i \in \mathbb{R}^p, y_i \in \{-1, 1\} \right\}_{i=1}^n \tag{7}$$

where y_i is either 0 or 1, and n is the number of training data. Each x_i is a p-dimensional vector having the feature quantity $\mathbb{R}$. Any hyper-plane can be written as:

$$w \cdot x - b = 0 \tag{8}$$

where, w is the vector to the hyper-plane. If the training data are linearly separable, then the hyper-plane can be described as:

$$w \cdot x - b = 1 \text{ and } w \cdot x - b = -1 \tag{9}$$

The distance between these two hyper-planes is $2/\|w\|$, and so the purpose is to minimize w. Therefore, the algorithm can be rewritten as:

$$\text{Minimize: } \|w\|, \text{ under the condition of } y_i(w \cdot x_i - b) \geq 1, \text{ for any } 1 \leq \text{i} \leq \text{n} \tag{10}$$

We can also reformulate the equation without changing the solution as:

$$\arg\min_{(w,b)} \frac{1}{2}\|w\|^2, \text{ under the condition of } y_i(w \cdot x_i - b) \geq 1, \text{ for any } 1 \leq \text{i} \leq \text{n} \tag{11}$$

The hyper-plane, or a set of hyper-planes, can be used as the separate lines in a classification. The SVM approach has recently been used in several financial applications (Martens et al. 2008; Tasi 2008; Li and Sun 2009; Chen et al. 2014; Yeh et al. 2010, 2014).

Class and regression tree (CART)

Classification and regression tree (CART) is a flexible method to describe how the variable Y is distributed after assigning the forecast vector X (Patil et al. 2012). It is able to classify huge amounts of data according to the division rule so as to identify valid data and thereby achieve ideal results (Kirkos et al. 2007a, b; Salehi and Fard 2013; Kim

and Upneja 2014; Marsala and Petturiti 2015). CART uses the binary tree to divide the forecast space into certain subsets on which the target variable distribution is continuously even. The "leaf" nodes correspond to different division areas that are determined by Splitting Rules relating to each internal node. By moving from the tree root to the leaf node, any forecast sample will be given only a leaf node.

This algorithm uses the GINI Index to determine in which attribute the branch should be generated. The building process of the model is to choose the attribute whose GINI index is a minimum after splitting. It can be described as:

$$GINI(T) = 1 - \sum_{i=1}^{m} P_i^2 \tag{12}$$

Let X be divided into n subsets, $\{T1, T2, \ldots Tn\}$. Among them, T_i's sample number is n_i. Thus, the Gini index divided according to property X is described as:

$$GINI(T) = 1 - \sum_{i=1}^{n} \frac{n_i}{n} GINI(T_i) \tag{13}$$

CART divides the property that leads a minimum value after the division.

Empirical analysis

Data collection and sampling

Research samples are drawn from GCD and NGCD firms in Taiwan from 2002 to 2013. 48 GCD firms are selected from all the listed companies of the Taiwan Economic Journal (TEJ) Data Bank. We adopt the 1-by-3 pair technique in order to match 144 NGCD firms. Thus, there are 192 firms in total that serve as our research sample of GCD and NGCD firms as shown in Table 1. Based on the indicators' selection in prior studies on GCD (Anandarajan and Anandarajan 1999; Behn et al. 2001; Kirkos et al. 2007a, b; Martens et al. 2008; Yeh et al. 2014), we prepare a set of 22 variables, as displayed in Table 2. These indicators are available in the TEJ database.

For the consideration of the number of samples, in order to avoid having too few samples in the test group and in order to improve test accuracy, we randomly gather 5 subsets from our original sample set and conduct fivefold cross validation.

Model development

This study begins by reducing the indicators using the LASSO screening method. The variables screened serve as the input variables for NN, CART and SVM. Next, the study carries out the model training and testing with every method. Finally, the study

Table 1 Samples

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	Total
GCD samples	20	2	4	4	4	1	4	2	2	1	2	2	48
NGCD samples	60	6	12	12	12	3	12	6	6	3	6	6	144

Table 2 Research variables

No.	Variable description/Definition or formula	Sources
X1	Total assets: Natural logarithm of total assets	Zhou et al. (2012), Chen et al. (2014), Yeh et al. (2014) and Chen and Lee (2015)
X2	Net sales: Natural logarithm of net sales	Tang and Firth (2011) and Chen et al. (2014)
X3	Current ratio: Current assets/Current liabilities	Lin (2009), Huang and Lu (2000), Sun et al. (2011), Zhou et al. (2012), Yeh et al. (2014), Chen and Lee (2015) and Chen et al. (2015)
X4	Debt ratio: Total liabilities/Total assets	Lin (2009), Huang and Lu (2000), Yeh et al. (2010), Jiang and Habib (2012), Chen et al. (2014, 2015), Yeh et al. (2014) and Chen and Lee (2015)
X5	Current assets: Natural logarithm of current assets	Korol (2013)
X6	Undistributed surplus: Natural logarithm of undistributed surplus	Chen and Lee (2015)
X7	Long term liabilities: Natural logarithm of long term liabilities	Korol (2013)
X8	Inventory: Natural logarithm of inventory	Salehi and Fard (2013)
X9	Total equity: Natural logarithm of total equity	Korol (2013)
X10	Total liabilities: Natural logarithm of total liabilities	Chen and Lee (2015)
X11	Net profit before tax: Income before tax	Chen et al. (2015)
X12	Operating cash flow: Cash flow from operating activities	Jiang and Habib (2012) and Chen et al. (2015)
X13	Accounts receivable turnover: Net sales/Average accounts receivable	Sun and Li (2008), Huang and Lu (2000), Yeh et al. (2010), Chen and Lee (2015) and Chen et al. (2015)
X14	Inventory turnover: Cost of goods sold/Average inventory	Zhou et al. (2012), Chen and Lee (2015) and Chen et al. (2015)
X15	Stockholding ratio of directors and supervisors: Number of stocks held by directors and supervisors/Total number of common stock outstanding	Chen and Lee (2015) and Chen et al. (2015)
X16	Big CPA firm or not (Big 4 in Taiwan): 1 for companies audited by BIG4, otherwise is 0	Jiang and Habib (2012), Yeh et al. (2014), Chen and Lee (2015) and Chen et al. (2015)
X17	Change CPA firm (CPA) or not: 1 is for change; 0 is for non-change	Anandarajan and Anandarajan (1999), Yeh et al. (2014) and Chen and Lee (2015)
X18	Current liabilities: Natural logarithm of current liabilities	Salehi and Fard (2013)
X19	Operating income: Natural logarithm of operating income	Salehi and Fard (2013) and Chen et al. (2015)
X20	Total assets turnover: Net Sales/Average total assets	Sun and Li (2008) and Sun et al. (2011)
X21	Earnings before interest and tax (EBIT)	Salehi and Fard (2013) and Chen et al. (2015)
X22	Return on assets (ROA): [Net income + interest expense × (1−tax rate)]/Average total assets	Martens et al. (2008), Lin (2009), Sun et al. (2011), Zhou et al. (2012), Jiang and Habib (2012) and Chen et al. (2015)

compares the merits and demerits of the classification ratio and provides relevant suggestions based on the analytic results.

Model construction is divided into three parts. The first part is replacement sampling; the second part is the LASSO feature selection; and the third part compares the test results of four kinds of classification models. The research process of this study is shown in Fig. 2.

Important variable screening

While constructing the classification model, many variables may be included, but not all of these variables are actually important. Therefore, unimportant variables need to be

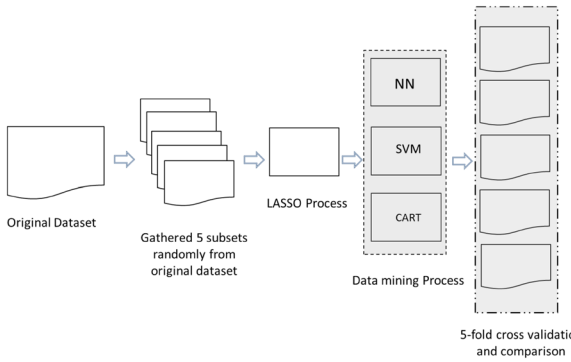

Fig. 2 Research process

eliminated in order to construct a simpler classification model. There is quite a number of ways to screen variables, of which the LASSO algorithm has shown excellent performance in reducing variables (Connor et al. 2015).

This study therefore adopts the suggestions of Connor et al. (2015) and screens the important indicators using the LASSO technique in order to retain only input variables with a significant influence. We employ the LASSO available in the SAS software to calculate the AIC values and coefficients of variable importance. The input variables of the study are screened using LASSO to acquire the results shown in Table 3 and Figs. 3, 4, 5, 6 and 7.

This study proposes a GCD prediction model for CPAs. Thus, the study adopts the indicators as input variables, which were selected in each screening process (Work-Groups 1–5). The important variables selected by using LASSO include: X4 (Debt ratio), X6 (Undistributed surplus), X20 (Total assets turnover), and X22 (Return on assets; ROA).

Table 3 LASSO variables' screening process

Steps	Work-G1 (AIC)	Work-G2[a] (AIC)	Work-G3 (AIC)	Work-G4[b] (AIC)	Work-G5 (AIC)
1	X4 (−77.5676)	X4 (−94.7118)	X4 (−66.0500)	X4 (−83.1760)	X4 (−71.2937)
2	X22 (−108.2326)	X6 (−93.3790)	X6 (−80.9976)	X22 (−83.9267)	X22 (−115.3547)
3	X11 (−116.1226)	X22 (−94.0645)	X22 (−79.4015)	X6 (−87.0297)	X11 (−125.4222)
4	X6 (−127.3604)	X19 (−93.0137)	X19 (−129.3612)	X20 (−85.2646)	X20 (−123.5628)
5	X20 (−146.4499)	X20 (−100.9320)	X13 (−134.4688)	X15 (−94.1284)	X6 (−124.3376)
6	X7 (−152.5126)	X15 (−101.0658)	X14 (−132.8479)	X11 (−95.2185)	X14 (−133.9785)
7	X5 (−152.5561)	X17 (−100.642)	X20 (−134.1510)	X14 (−107.4634)	X16 (−134.0137)
8		X14 (−104.7244)	X17 (−136.4395)	X1 (−120.0362)	
9		X11 (−102.8433)	X16 (−142.2861)	X9 (−120.4143)	
10		X13 (−107.1809)			
11		X5 (−107.8717)			
12		X12 (−116.8996)			
13		X16 (−124.2823)			

[a] X9 effect entered at step, AIC value is −104.7244, removed at step 13, AIC value form −107.8717 decease to −115.5186

[b] X21 effect entered at step 5, AIC value is −93.7699, removed at step 9, AIC value form −107.4634 decease to −112.5140

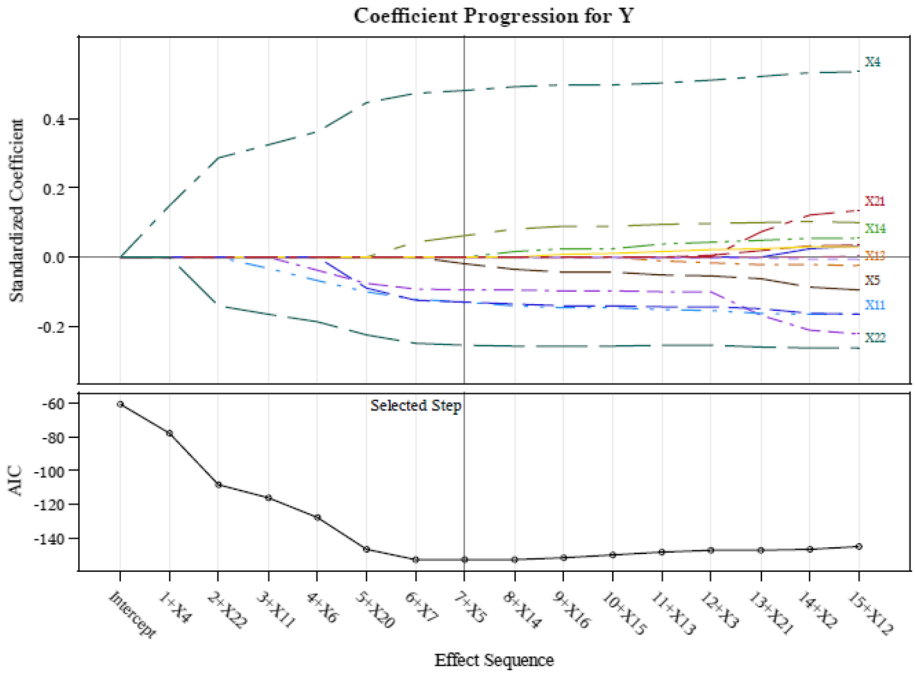

Fig. 3 LASSO variables screening process Work-Group 1

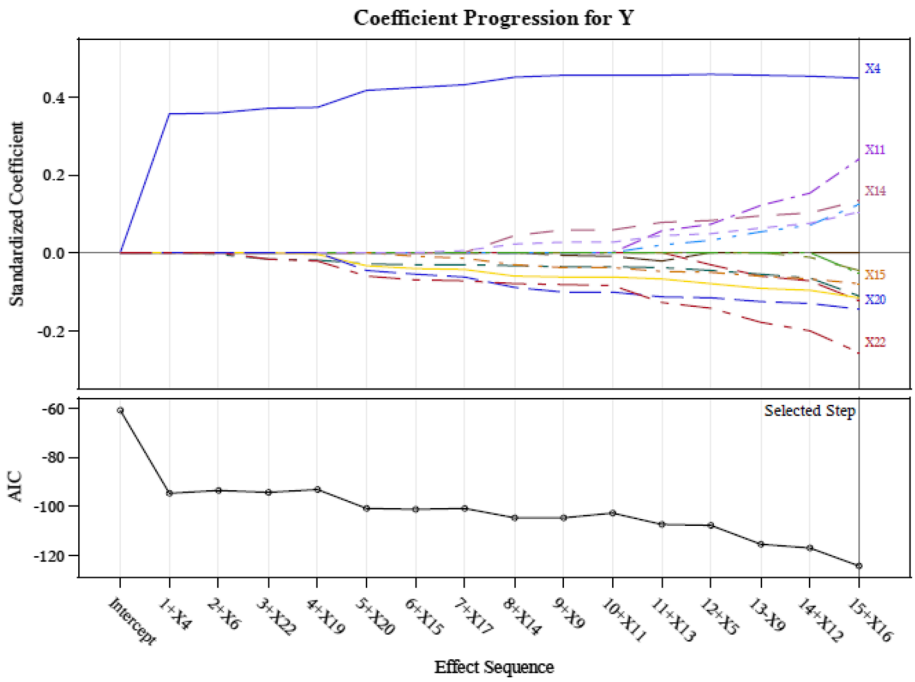

Fig. 4 LASSO variables screening process Work-Group 2

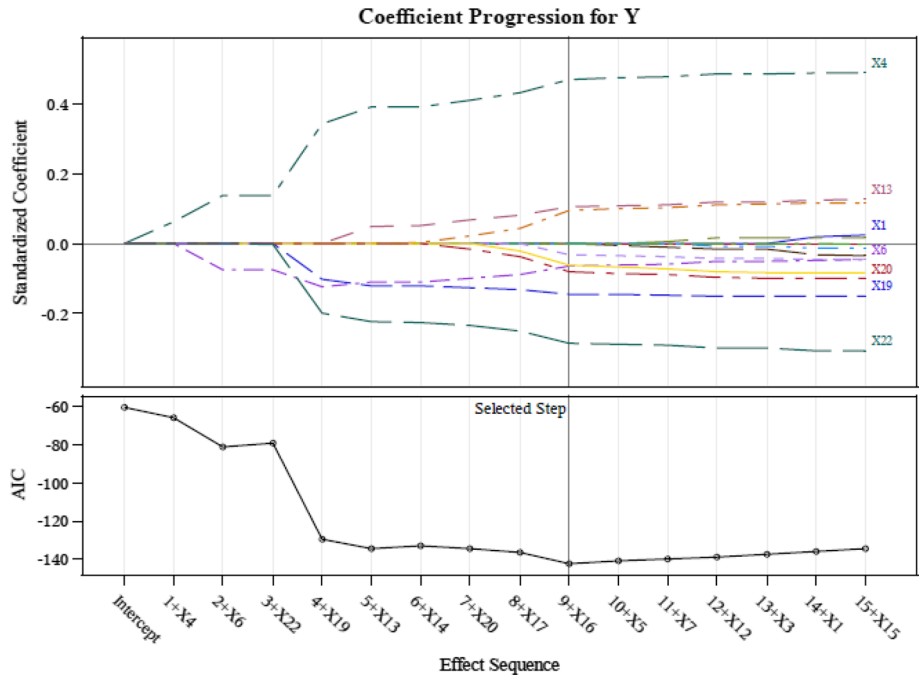

Fig. 5 LASSO variables screening process Work-Group 3

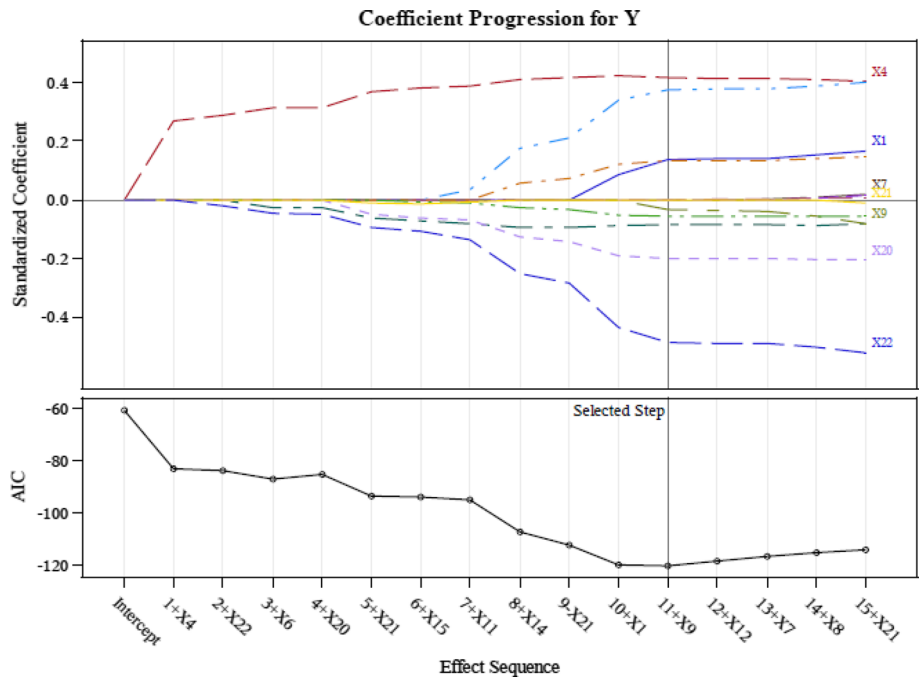

Fig. 6 LASSO variables screening process Work-Group 4

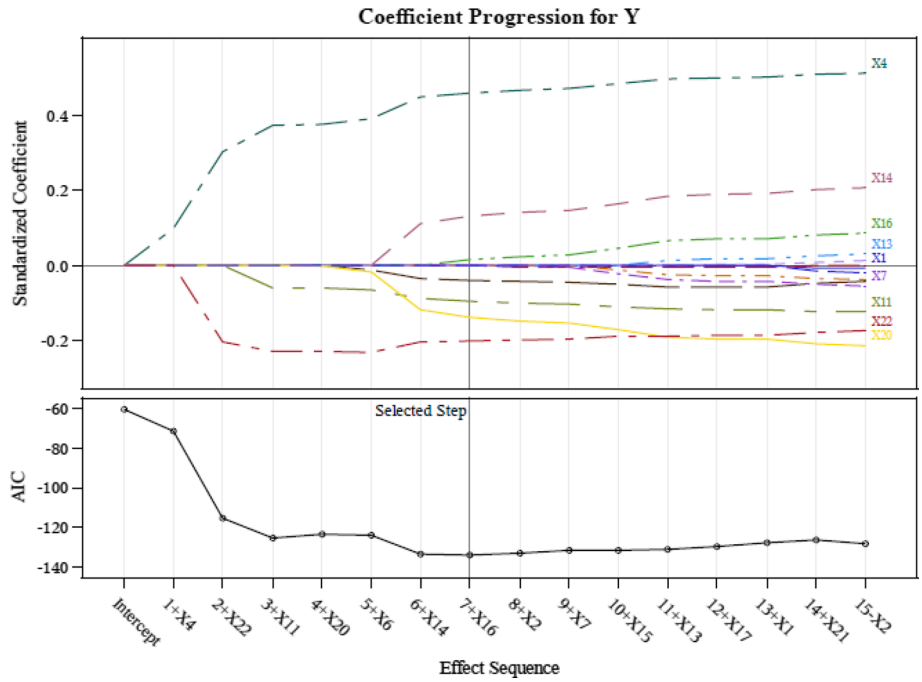

Fig. 7 LASSO variables screening process Work-Group 5

X4 (Debt ratio: Total liabilities/Total assets) is an important measure of the debt ratio and capital structure of a company. Generally, capital is sourced from stockholders or external financing. Financing has a leverage that can increase the return on investment. Moreover, interest costs are not taxed, and thus financing has numerous advantages, but if debt is high, then financial leverage may increase risk. If a firm's operations are not as good as expected, then bankruptcy may occur. X6 (Undistributed surplus) is net income after withdrawal of legal and special surplus and can be used to pay cash dividends, expansion, or R&D. X20 (Total assets turnover: Net Sales/Average total assets) is an important measure to evaluate the operation quality of corporate assets and utilization efficiency. The greater the turnover rate is, the faster the turnover of total assets, and the stronger the sales ability. X22 (Return on assets (ROA): [Net income + interest expense × (1 − tax rate)]/Average total assets) shows the percentage of how profitable a company's assets are in generating revenue.

This study subsequently takes the 4 variables above as new input predictors in order to construct a prediction/classification model. The descriptive statistics and correlation of input variables are shown as Tables 4 and 5.

Classification model

This study employs IBM SPSS modeler 14.0 to build classification models NN, CART, and SVM. The cross-validation results of the training and testing subsets are shown as Tables 6, 7 and 8.

Table 4 Descriptive statistics of input variables

Variable		N	Mean	SD	Min	Max
X4	Debt ratio	192	51.0965	21.6263	4.8700	101.9700
X6	Undistributed surplus	192	−346,749.52	2,210,187.98	−22,801,544.00	5,561,297.0000
X20	Total assets turnover	192	0.8593	0.6895	0.0300	4.8400
X22	Return on assets (ROA)	192	−0.0756	0.2762	−2.0997	0.3695

Table 5 Correlation of input variables

Input variable		X4	X6	X20	X22
X4	Debt ratio	1	–	–	–
X6	Undistributed surplus	−0.3137	1	–	–
		<0.0001			
X20	Total assets turnover	0.0048	0.2430	1	–
		0.9478	0.0007		
X22	Return on assets (ROA)	−0.2752	0.2146	0.1941	1
		0.0001	0.0028	0.0070	

Table 6 LASSO–NN model—the fivefold cross validation results

Subset	Training set					Testing set				
	Predicted group		Hit ratio (%)	Type I error (%)	Type II error (%)	Predicted group		Hit ratio (%)	Type I error (%)	Type II error (%)
1	71	1	98.96	1.39	0.00	70	2	94.79	2.78	12.50
	0	24				3	21			
2	70	2	90.62	2.78	29.17	60	12	85.42	16.67	8.33
	7	17				2	22			
3	69	3	92.71	4.17	4.17	64	8	90.62	11.11	4.17
	1	23				1	23			
4	60	12	87.50	16.67	12.50	59	13	85.42	18.06	4.17
	3	21				1	23			
5	70	2	96.88	2.78	4.17	63	9	88.54	12.50	8.33
	1	23				2	22			
Avg.			93.33	5.56	10.00			88.96	12.22	7.50

LASSO–NN model

The NN model is set as follow: (1) model type is set at Multilayer Perceptron (MLP), one hidden layer, and maximum training cycles stop at 250 times. The LASSO–NN model classification results are shown as Table 6.

On average, 9 of the 72 NGCD materials are incorrectly classified, and the Type I error rate is 12.22 %. In addition, 22 of the 24 GCD materials are correctly classified, while the remaining 2 GCD materials are incorrectly classified in NGCD. The Type II error is 7.50 %. The weight of each node and importance of variables are shown as Figs. 8 and 9.

Table 7 LASSO–CART model—the fivefold cross validation results

Subset	Training set				Testing set			
	Predicted group	Hit ratio (%)	Type I error (%)	Type II error (%)	Predicted group	Hit ratio (%)	Type I error (%)	Type II error (%)
1	66 6 0 24	93.75	8.33	0.00	68 4 2 22	93.75	5.56	8.33
2	70 2 4 20	93.75	2.78	16.67	57 15 4 20	86.46	20.83	16.67
3	67 5 2 22	92.71	6.94	8.33	65 7 5 19	90.62	9.72	20.83
4	69 3 4 20	92.71	4.17	16.67	60 12 3 21	83.33	16.67	12.50
5	72 0 5 19	94.79	0.00	20.83	61 11 3 21	89.58	15.28	12.50
Avg.		93.54	4.44	12.50		88.75	13.61	14.17

Table 8 LASSO–SVM model—the fivefold cross validation results

Subset	Training set				Testing set			
	Predicted group	Hit ratio (%)	Type I error (%)	Type II error (%)	Predicted group	Hit ratio (%)	Type I error (%)	Type II error (%)
1	71 1 2 22	96.88	1.39	8.33	66 6 2 22	91.67	8.33	8.33
2	70 2 7 17	90.62	2.78	29.17	66 6 4 20	89.58	8.33	16.67
3	71 1 6 18	92.71	1.39	25.00	66 6 5 19	88.54	8.33	20.83
4	68 4 8 16	87.50	5.56	33.33	62 10 3 21	86.46	13.89	12.50
5	72 0 3 21	96.88	0.00	12.50	70 2 5 19	92.71	2.78	20.83
Avg.		92.92	2.22	21.67		89.79	10.00	15.83

LASSO–CART model

This study constructs the LASSO–CART model, sets maximum depth at 5, and adopts the Gini index as an impurity measure for categorical targets. The forecast results of the LASSO–CART prediction model are shown in Table 7. On average, 62 of the 72 NGCD materials are correctly classified, while 10 of them are incorrectly classified in GCD, for a Type I error of 13.61 %. On the other hand, 20 of the 24 GCD materials are correctly classified, with the remaining 2 GCD materials incorrectly classified in NGCD. The Type II error is 14.17 %.

LASSO–SVM model

In terms of the LASSO–SVM model, the kernel type is set at "Linear", the stopping criteria is set at 1.0E−3, and the regularization parameter is set at 10 and 0.1 of the regression precision.

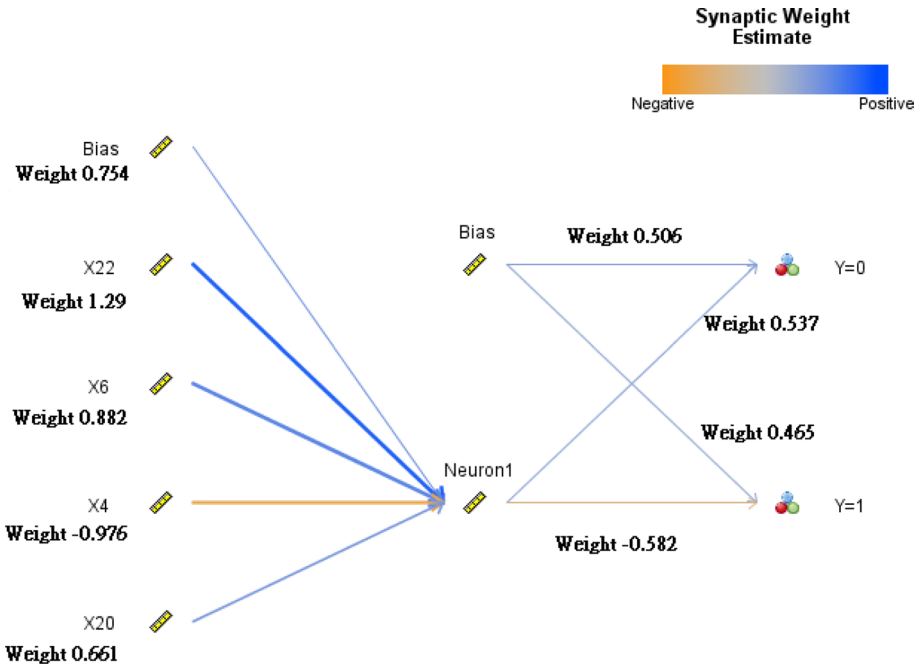

Fig. 8 Weight of each node of the NN model

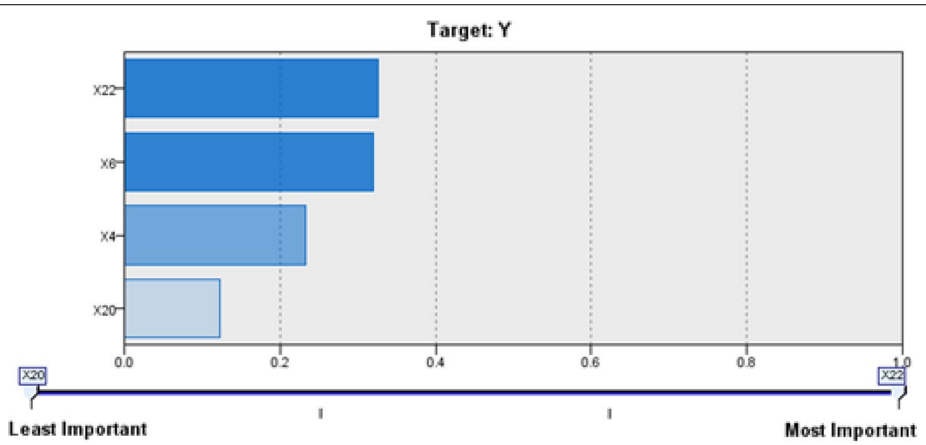

Fig. 9 Importance of variables

The LASSO–SVM classification results are shown in Table 8. On average, 66 of the 72 NGCD materials are correctly classified, while 6 of them are incorrectly classified in GCD. The Type I error is 10.00 %. In addition, 20 of the 24 GCD materials are correctly classified, with the remaining 4 GCD materials incorrectly classified in NGCD. The Type II error is 15.83 %.

Model comparison and statistical test

According to the empirical results (Tables 6, 7, 8), the prediction accuracy of the LASSO–NN model is 88.96 % (Type I error rate is 12.22 %; Type II error rate is 7.50 %), the prediction accuracy of the LASSO–CART model is 88.75 % (Type I error rate is 13.61 %; Type II error rate is 14.17 %), and the prediction accuracy of the LASSO–SVM model is 89.79 % (Type I error rate is 10.00 %; Type II error rate is 15.83 %). Our comparison follows that of Kirkos et al. (2007a, b), Tasi and Huang (2010) and Chen et al. (2014). We not only focus on the hit ratio of the models, but also consider the Type I error and Type II error rates.

Unlike past works, which typically use Type I errors to judge the performance of a forecasting model, GCP studies prefer to use Type II errors to determine the performance of forecasting models. In order to confirm the significant difference between prediction models, this study uses the Wilcoxon two-sample test and the Kruskal–Wallis test, with the results shown in Table 9. The test results reveal a significant difference among the LASSO–NN, LASSO–CART, LASSO–NN, and LASSO–SVM prediction models.

Conclusions

Certified public accountants (CPAs) and auditors check firms' financial statements and issue their audit opinions and audit reports. These audit opinions and audit reports are very important for enterprises, stakeholders, and financial markets, especially investors. Thus, it is necessary to establish more accurate going concern doubt prediction models. The purpose of this study is to set up rigorous and reliable going concern doubt prediction models for auditors. This study applies the least absolute shrinkage and selection operator (LASSO) and data mining techniques (NN, CART, and SVM) to establish the prediction models.

According to the empirical results, the prediction accuracy is 88.96 % for the LASSO–NN model, is 88.75 % for the LASSO–CART model, and is 89.79 % for the LASSO–SVM model. This study uses LASSO to select important variables, which include: X4 (Debt ratio), X6 (Undistributed surplus), X20 (Total assets turnover), and X22 (Return on assets; ROA). As such, a firm's top management, CPAs, and auditors all should pay close attention to them.

Type I errors may not have serious consequences when compared to Type II errors. If the auditor wrongly classifies a GC firm as healthy, then he/she can be sued. If an auditor issues a wrong audit report due to his/her misjudgment, then this will affect not only the

Table 9 Statistical tests

Statistical test method	Statistical test	NN–CART	NN–SVM
Wilcoxon test	Z	−1.9335	−2.0280
	one-sided pr <Z	0.0266	0.2130
	two-sided pr <|Z|	0.0532*	0.0426**
Kruskal–Wallis test	Chi square	4.1654	4.5570
	DF	1	1
	Pr >Chi square	0.0413**	0.0328**

* Significant at P < 0.1; ** significant at P < 0.05, *** significant at P < 0.01

enterprise and stakeholders, but also many investors. Moreover, the CPA may be sued. The costs for Type II errors are thus rather severe. We have developed three GCD prediction models. In the LASSO–NN model, the Type I error rate is 12.22 % and the Type II error rate is 7.50 %; in the LASSO–CART model, the Type I error rate is 13.61 % and the Type II error rate is 14.17 %; and in the LASSO–SVM model, the Type I error rate is 10.00 % and the Type II error rate is 15.83 %. These error rates are all lower than 20 %, especially in the LASSO–NN model where the Type II error rate is only 7.50 %. This is a key contribution of this paper.

Finally, the empirical results of this study can provide a reference for enterprises' top management, CPAs, auditors, and future studies.

Limitations

There are several limitations in this study. 1. The size of the financial market in Taiwan is not as big when compared to China, the U.S., UK, EU, Japan, etc.); 2. The Taiwan government has strict control over the listed companies and the financial market. Thus, GCD listed companies are fewer. 3. If the GCD prediction models are used in countries other than Taiwan, then the GCD indicators (variables) should be measured according to national or economically regional audit laws and regulations and financial practice.

Authors' contributions

YG and DC made substantial contributions to the concept and design of the present study. ZS made substantial contributions to acquisition and interpretation of the data and research methods. All authors read and approved the final manuscript.

Author details

[1] Department of Business Administration, National Taipei University, No. 67, Section 3, Ming-Shen East Road, Taipei City 10478, Taiwan. [2] Department of Accounting, Chinese Culture University, No. 55, Hwa-Kang Road, Yang-Ming-Shan, Taipei City 11114, Taiwan.

Acknowledgements

The authors thank the editor-in-chief, editors, and the anonymous reviewers of SpringerPlus for their insightful comments, which have helped to improve the quality of this paper.

Competing interests

The authors declare that they have no competing interests.

Received: 14 November 2015 Accepted: 19 April 2016
Published online: 27 April 2016

References

Anandarajan M, Anandarajan A (1999) Comparison of machine learning techniques with a qualitative response model for auditors' going concern reporting. Expert Syst Appl 16(4):385–392

Behn BK, Kaplan SE, Krumwiede KP (2001) Further evidence on the auditor's going-concern report: the influence of management plans. Audit J Pract Theory 20(1):13–29

Benston G, Hartgraves AL (2002) Enron: what happened and what we can learn from it. J Account Public Policy 21(2):105–127

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Haussler D (ed) Proceedings of the annual conference on computational learning theory. ACM Press, Pittsburgh, PA, pp 144–152

Brabazon A, Keenan B (2004) A hybrid genetic model for the prediction of corporate failure. CMS 1(3):293–310

Carcello JV, Neal TL (2000) Audit committee composition and auditor reporting. Account Rev 75(4):453–467

Chen KCW, Church BK (1992) Default on debt obligations and the issuance of going-concern opinions. J Pract Theory 11(2):30–50

Chen KCW, Church BK (1996) Going concern opinions and the market's reaction to bankruptcy filings. Account Rev 71(1):117–128

Chen S, Lee J (2015) Going concern prediction using data mining. ICIC Express Lett Part B Appl 6(12):3311–3317

Chen S, Goo JYJ, Shen ZD (2014) A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements. Sci World J 2014:968712. doi:10.1155/2014/968712

Chen FH, Chi DJ, Wang YC (2015) Detecting biotechnology industry's earnings management using Bayesian network, principal component analysis, back propagation neural network, and decision tree. Econ Model 46:1–10

Colombani C, Legarra A, Fritz S, Guillaume F, Croiseau P, Ducrocq V (2013) Application of Bayesian least absolute shrinkage and selection operator (LASSO) and Bayes Cπ methods for genomic selection in frenchholstein and montbeliarde breeds. J Dairy Sci 96(1):575–591
Connor P, Hollensen P, Krigolson O, Trappenberg T (2015) A biological mechanism for Bayesian feature selection: weight decay and raising the LASSO. Neural Netw 67:121–130
Cornier D, Magnan M, Morard B (1995) The auditor's consideration of the going concern assumption: a diagnostic model. J Account Audit Finance 10(2):201–221
Foster B, Ward T, Woodroof J (1998) An analysis of the usefulness of debt defaults and going concern opinions in bankruptcy risk assessment. J Account Audit Finance 13(3):351–371
Gaganis C, Pasiouras F, Doumpos M (2007) Probabilistic neural networks for the identification of qualified audit opinions. Expert Syst Appl 32:114–124
Geiger MA, Rama DV (2003) Audit fees, non-audit fees, and auditor reporting on stressed companies. Audit J Pract Theory 22(2):53–69
Harrell FE (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, New York
Huang CL, Lu SC (2000) A study of company financial distress warning model-constructing with financial and non financial factors. J Contemp Account 1(1):19–40
Jiang H, Habib A (2012) Split-share reform and earnings management: evidence from China. Adv Account Inc Adv Int Account 28:120–127
Kim SY, Upneja A (2014) Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Econ Model 36:354–362
Kirkos S, Spathis C, Manolopoulos Y (2007a) Data mining techniques for the detection of fraudulent financial statements. Expert Syst Appl 32(4):995–1003
Kirkos E, Spathis C, Nanopoulos A, Manolopoulos Y (2007b) Identifying qualified auditors' opinions: a data mining approach. J Emerg Technol Account 4(1):183–197
Koh HC, Low CK (2004) Going concern prediction using data mining techniques. Manag Audit J 19(3):462–476
Korol T (2013) Early warning models against bankruptcy risk for central European and Latin American enterprises. Econ Model 31:22–30
Lenard MJ, Alam P, Madey GR (1995) The application of neural networks and a qualitative response model to the auditor's going concern uncertainty decision. Decis Sci 26(2):209–227
Li H, Sun J (2009) Predicting business failure using multiple case-based reasoning combined with support vector machine. Expert Syst Appl 36:10085–10096
Lin TH (2009) A cross model study of corporate financial distress prediction in Taiwan: multiple discriminant analysis, logit, probit and neural networks models. Neurocomputing 72:3507–3516
Marsala C, Petturiti D (2015) Rank discrimination measures for enforcing monotonicity in decision tree induction. Inf Sci 291(10):143–171
Martens D, Bruyneseels L, Baesens B, Willekens M, Vanthienen J (2008) Predicting going concern opinion with data mining. Decis Support Syst 45(4):765–777
Mokhatab RF, Manzari SM, Bostanian S (2011) Financial health prediction models using artificial neural networks, genetic algorithm and multivariate discriminant analysis: Iranian evidence. Expert Syst Appl 38(8):10210–10217
Mutchler JF, Hopwood WS, McKeown JC (1997) The influence of contrary information and mitigating factors on audit opinion decisions on bankrupt companies. J Account Res 35(2):295–310
O'Leary DE (1998) Using neural network to predict corporate failure. Int J Intell Syst Account Finance Manag 7(3):187–197
Patil A, Dyavaiah M, Joseph F, Rooney JP, Chan CT, Dedon PC, Begley TJ (2012) Increased tRNA modification and gene-specific codon usage regulate cell cycle progression during the DNA damage response. Cell Cycle 11(19):3656–3665
Salehi M, Fard FZ (2013) Data mining approach to prediction of going concern using classification and regression tree (CART). Glob J Manag Bus Res Account Audit 13(3):25–29
Sun J, Li H (2008) Data mining method for listed companies' financial distress prediction. Knowl Based Syst 21:1–5
Sun J, He KY, Li H (2011) SFFS-PC-NN optimized by genetic algorithm for dynamic prediction of financial distress with longitudinal data streams. Knowl Based Syst 24:1013–1023
Tang T, Firth M (2011) Can book–tax differences capture earnings management and tax management? Empirical evidence from China. Int J Account 46:175–204
Tasi CF (2008) Financial decision support using neural networks and support vector machines. Expert Syst 25(4):380–393
Tasi BH, Huang YP (2010) Alternative financial distress prediction models. J Contemp Account 11(1):51–78
Tasi CF, Wu JW (2008) Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Syst Appl 34:2639–2649
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc SerB (Methodol) 58(1):267–288
Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J R Stat Soc Ser B (Methodol) 67(1):91–108
Toiviainen P, Alluri V, Brattico E, Wallentin M, Vuust P (2014) Capturing the musical brain with Lasso: dynamic decoding of musical features from fMRI data. Neuroimage 88:170–180
Yamada M, Jitkrittum W, Sigal L, Xing EP, Sugiyama M (2014) High-dimensional feature selection by feature-wise kernelized lasso. Neural Comput 26(1):185–207
Yeh CC, Chi DJ, Hsu MF (2010) A hybrid approach of DEA, rough set and support vector machines for business failure prediction. Expert Syst Appl 37(2):1535–1541
Yeh CC, Chi DJ, Lin YR (2014) Going-concern prediction using hybrid random forests and rough set approach. Inf Sci 254:98–110
Zhou L, Lai KK, Yen J (2012) Empirical models based on features ranking techniques for corporate financial distress prediction. Comput Math Appl 64:2484–2496