Comparison of Single Classifier Models for Predicting Long-term Business Failure of Construction Companies Using Finance-based Definition of the Failure

H. Choi, H. Sung, H. Cho, S. Lee, H. Son, and C. Kim*

Department of Architectural Engineering, Chung-Ang University, South Korea
E-mail: vinaj516@gmail.com, gusdn7543@gmail.com, hyukmanjo@gmail.com, leesungwook@cau.ac.kr, hjson0908@cau.ac.kr, changwan@cau.ac.kr (*corresponding author)

Abstract

This study compares the performance of six classifier models (ANN, KNN, C4.5, SVM, LR, and NB) for predicting the business failure of construction companies after three years from 2010 to 2012. Although previous studies have explored numerous business failure prediction models for construction companies, these models have focused on short-term failure, defined as failure occurring within one year, and have defined business failure based on companies' experiencing serious legal events, including bankruptcy, delisting, and default. However, the construction industry is typically characterized by projects with longer durations, usually exceeding one year. This implies that previous short-term models cannot predict the business failure of construction companies until the end of particular projects. Moreover, this problem is compounded by the fact that legal events can involve lengthy proceedings, which are often initiated much later than the actual moment of the business failure. Therefore, in this study, six classifier models will be used to predict the business failure of the construction companies within three years using a finance-based definition of failure. The results show that all six models' performances noticeably decrease when they predict more than one year. These results demonstrate that previous short-term prediction models with outstanding performance cannot be practical in predicting the long-term business failure of construction companies.

Keywords – Data Mining; Machine Learning; Project Management; Business failure; Prediction model;

1 Introduction

Business failure for a construction company during a construction project can lead to a series of failures for subcontractors, which can also result in serious socio-economic problems. Compared to businesses in other industries, construction companies are more vulnerable to business failure as a result of various specificities [1], including the uniqueness of the projects, the long duration of the projects, and the industry's sensitivity to economic cycles [2]. For example, the construction industry had the highest rate of bankruptcy among all industries in Korea between 1998 and 2015 [3]. Due to this vulnerability, the ability to predict a potential business failure for a construction company in the early stages of a construction project can be a tool of critical importance to a variety of stakeholders, including the project owner, investors, creditors, and contractors.

Business failure prediction models for construction companies based on financial ratios have been proposed since 1970. These models are based on statistical techniques, such as multivariate discriminant analysis (MDA) [4] and logistic regression (LR) [5], and, more recently, data mining techniques [2,6–9]. Most of the previous studies have focused on short-term (within one year) predictions, even though construction projects typically have a longer duration. Thus, these models cannot predict the construction companies' financial risk during the entire project period. Also, these short-term models may suffer from an absence of data because, in many cases, the publication of the annual accounts of the failing companies can be delayed [10]. Moreover, these previous studies defined the business failure based on highly visible legal events from a sample construction company, including bankruptcy, delisting, and default. However, legal events of this type are typically the culmination of a lengthy legal process, meaning their occurrence can be much later than the moment of actual failure [10,11]. To address this issue, this study compared single classifier models for predicting the business failure of a construction company within three years in order to cover the relatively long duration of the projects. Moreover, this study uses a finance-based definition of business failure regardless of legal events.

The rest of the paper is organized as follows. Section 2 reviews the literature on the prediction term, the definition of the business failure, and the techniques used in their models. Section 3 describes the collected data and the selected variables. Section 4 describes the pre-processing of the data and the single classifier model employed in this study, including 6 single classifiers (support vector machine [SVM], artificial neural networks [ANN], commercial version 4.5 [C4.5], naive Bayes [NB], LR, k-nearest neighbor [KNN]). Section 5 presents the results of experiments designed to evaluate and compare the performances of the six single classifiers in terms of the area under the receiver operating characteristic curve (AUC). Finally, conclusions are presented in Section 6.

2 Literature Review

Recently, business failure prediction models for construction companies based on data mining techniques [1,2] have been found to outperform the traditional statistical models based on MDA [4] and LR [5]. Therefore, the models for predicting business failure using data mining is actively proposed [1,2,6–9]. However, previous studies have focused on short-term (within one year) prediction and have defined business failure based on legal events.

Tserng et al. [2] proposed a model for predicting the default of construction companies within one year. They used the data of the year right before default as the data of default companies. Default was defined as delisting due to bankruptcy, liquidation, or poor performance. They compared an enforced SVM (ESVM) model with the LR model. The results show that ESVM outperforms the traditional LR model. Chen and Hoang [6] proposed a model for predicting financial distress in a quarter of construction companies, using data from quarterly financial reports. They defined financial distress as bankruptcy, delisting, bounding, bailouts, or major organizational restructuring. Their model is based on self-organizing feature map optimization and fuzzy- and hyper-rectangular composite neural networks. Horta and Camanho [1] proposed a model for predicting company failure within one year in construction industries. Their proposed model, based on SVM, outperforms the LR model in predicting failure. Heo and Yang [7] proposed a model for predicting bankruptcy within one year. They used the data one year prior to bankruptcy as the data of the bankrupt companies. They defined bankruptcy as workout, receivership, or bankruptcy. Their model was based on adaptive boosting (AdaBoost). The proposed AdaBoost model was compared with the ANN, SVM, decision tree (D-Tree), and Z-Score model. The results showed that the proposed AdaBoost model outperforms other models. Cheng et al. [8] and Tserng et al. [9] proposed a model for predicting the default of construction companies within one year. They defined default as delisting due to bankruptcy, liquidation, or poor performance, as in Tserng et al. [2]. Their models are based on least squares SVM and grey system theory, respectively. As can be seen, the relatively long duration of construction projects and the financial risk construction companies could experience before the occurrence of legal events have not been sufficiently considered.

3 Data and Variable Selection

From Korean information service value [12], this study obtained data derived from financial statements of construction companies with Korean standard industrial classification codes 41 and 42 covering 2007 to 2012. The data for years 2007, 2008, and 2009 were used to predict business failure after one year (2010), two years (2011), and three years (2012), respectively. The data were organized into three datasets, which are a 2010 dataset for predicting after one year, a 2011 dataset for predicting after two years, and a 2012 dataset for predicting after three years. This study excluded companies that did not provide complete data during the sample period to guarantee the completeness of the data. Companies with total assets less than 12 billion won or with total assets and total liabilities less than 7 billion won were also excluded in order to ensure the reliability of the data, which is one of the conditions of the external auditing company. Finally, the financial information of 385 construction companies was used as the final data.

This study used a finance-based definition of business failure to classify companies into normal and failed companies regardless of legal events. The finance-based definition of failure has been used to detect failure regardless of the legal consequence used in previous studies [11, 13-15]. Previous studies [e.g. 14–17] employed special treatment (ST) regulation to define business failure. ST regulation is the early warning system to identify abnormalities in Chinese-listed companies’ financial status [17]. Ding et al. [18] showed that the companies fell into business failure after receiving the ST and showed some signs of potential failure prior to receiving it. Thus, predicting ST can be a financial early warning system for possible future business failure. Therefore, this study utilized the ST regulation to define business failure based on Li and Sun [14] and Sun and Li [15]. Therefore, companies having negative net income for two consecutive years were classified as failed companies. As a result, the numbers of failed companies in the three years’ datasets were 28, 41, and 49, as shown in Table 1.

Twenty-one financial ratios were selected as input variables based on a review of the previous studies [1,2,6–9]. Following Horta and Camanho [1] and Tserng et al [2], financial ratios were classified into four categories: activity, leverage, liquidity, and profitability. Table 2 provides definitions of these financial ratios, which cover a wide range of financial characteristics and performances [2], according to Cheng et al. [8].

Table 1. Distribution of the data

Dataset	Normal companies	Failed companies
2010	357	28
2011	344	41
2012	336	49

4 Methodology

The methodology framework is shown in Figure 1 and the process is as follows. First, an oversampling technique based on the synthetic minority oversampling technique (SMOTE) was employed to obtain a balanced dataset in which the number of normal and failed companies are equivalent. Then, six classification models were applied and their prediction accuracies were compared: SVM, ANN, C4.5, NB, LR, and KNN. At the same time, the parameter optimization with grid search was executed for the six classifiers. Finally, we evaluated the models with an AUC value using 10-fold cross validation. This study employed the algorithms from Weka release 3.8.1 [19] in all experiments.

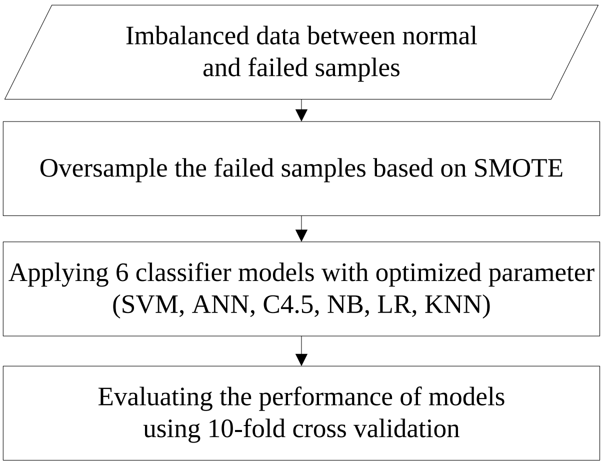
Imbalanced data between normal and failed samples
Oversample the failed samples based on SMOTE
Applying 6 classifier models with optimized parameter (SVM, ANN, C4.5, NB, LR, KNN)
Evaluating the performance of models using 10-fold cross validation

Figure 1. The Methodology Framework

Table 2. Financial ratios and definitions

Category	Variable	Definition
Activity	Accounts payable turnover	Sales / Average payable
	Accounts receivable turnover	Sales / Average receivables
	Current assets turnover	Sales / Current assets
	Fixed assets to net worth	(Total assets - Current assets) / Shareholders’ equity
	Quality of inventory	Cost of sales / Average inventories
	Revenues to fixed assets	Sales / (Total assets - Current assets)
	Revenues to net working capital	Sales / (Current assets - Current liabilities)
	Sales to net worth	Sales / Shareholders’ equity
	Turnover of total assets	Sales / Total assets
Leverage	Debt ratio	Total liabilities / Total assets
	Retained earnings to sales	Retained earnings / Sales
	Times interest earned	Earnings before interest and taxes / Interest expense
	Total liabilities to net worth	Total liabilities / Shareholders’ equity
Liquidity	Current assets to net assets	Current assets / (Total assets - Current liabilities)
	Current ratio	Current assets / Current liabilities
	Net working capital to total assets	(Current assets - Current liabilities) / Total assets
	Quick ratio	(Current asset - Inventories) / Current liabilities
Profitability	Profits to net working capital	Net income / (Current assets - Current liabilities)
	Return on assets	Net income / Total assets
	Return on equity	Net income / Shareholders' equity
	Return on sales	Net income / Sales

4.1 Pre-processing for Dealing with the Imbalanced Data Problem

As shown in Table 1, this study used three datasets with a population consisting of imbalanced data. However, the imbalanced data can distort the real performance of the prediction model. In other words, a model based on imbalanced data can yield high overall predictive accuracy driven by the majority class but with poor accuracy for the minority class [20]. Therefore, this study employed the over-sampling method SMOTE to handle this problem. SMOTE generates synthetic samples to oversample a minority class without data replication [21]. Therefore, SMOTE can avoid the over-fitting problem which can be caused from data replication [22–24]. In SMOTE, k nearest samples to the original samples are selected. This study set the value of k to 5, which is recommended by Chawla et al. [25]. Then, a new sample is generated by adding the average of the distances, multiplied by an arbitrary number from 0 to 1, between the selected k samples and the original sample to the original sample [26]. This process was repeated until the sample numbers of the minority class and the majority classes were balanced.

4.2 Six Single Classifiers with Optimized Parameters Using Grid Search

This study considered six different classifiers: SVM, ANN, C4.5, NB, LR, and KNN. The parameter values of each classifier were optimized using a grid search for three datasets as Table 3. The basic concept of a grid search involves applying various parameter values and choosing the one with the best cross-validation performance. In this study, the grid search was conducted using 10-fold cross-validation. Among the applied parameter values, the parameter of each classifier for each dataset are set to the parameter value with the best classification performance in 10-fold cross validation. Finally, the optimized parameters used in this study are summarized in Table 3.

4.3 Performance Evaluation Measures

This study employed an AUC measure to compare six classifier models, which is commonly used in evaluating business failure models for construction companies [1,2,8]. AUC is the most commonly employed as a summary statistic for the quality of the rankings [2]. This study carried out 10-fold cross validation to evaluate the classification performance of the six classifier models. This method is known to minimize bias and variance compared to all other validation methods [33,34]. In 10-fold cross validation, the training dataset was divided into ten subsets. Nine subsets were then used to train the model, and the remaining subset was retained to assess its performance. The performances of the classifiers were computed by averaging the performance of each of the ten subsets.

5 Experimental Results and Discussion

The performances of the six classifier models are summarized in Table 4. The results were generated by applying 10-fold cross validation by six classifiers to the three datasets for 2010, 2011, and 2012, respectively.

As shown in Table 4, the KNN model had the highest prediction performance among the six models. The ANN followed the KNN model. However, the C4.5 model has better performance in the case of predicting for 2012. This result shows that although the model has shown better performance results in the short-term prediction of business failure, the model may not show the best performance in long-term predictions. Furthermore, the results show that the performance of all six classifier models noticeably decreases when the models are used to predict more than one year ahead. This result indicates that the long-term prediction of business failure is much more difficult than short-term prediction. Therefore, although the short-term prediction models proposed in most previous studies have outstanding performance in short-term prediction, the model may not be useful for long-term prediction covering a relatively long project period in the construction industry.

Table 3. Grid search applied in this study

Classifiers	Parameters	Grid space	Optimized parameters for each dataset		
			2010	2011	2012
SVM [27,28]	Penalty parameter C	$(2^1, 2^2, \ldots 2^{15})$	16	8	8
	Kernel parameter γ	$(2^{-15}, 2^{-14}, \ldots 2^3)$	8	8	8
ANN [29]	Learning rate μ	(0.1,0.2,…,0.9)	0.1	0.7	0.7
	Momentum α	(0.1,0.2,…,0.9)	0.7	0.1	0.1
C4.5 [30,31]	Confidence Factor	(0,0.05,…,1)	1	1	1
	Minimum number of samples per leaf	(5,10,…,150)	5	15	10
KNN [32]	The number of closest neighbors k	(1,2,…,50)	9	7	4

Table 4. The performance(AUC) of the six models

Method	Prediction Year		
	2010	2011	2012
SVM	0.923	0.860	0.858
ANN	0.954	0.907	0.860
C4.5	0.927	0.876	0.862
NB	0.901	0.830	0.753
LR	0.912	0.844	0.794
KNN	0.955	0.903	0.894

6 Conclusion

This study presents a comparison of the prediction performances of six single classifier models for predicting the long-term business failure of 385 Korean construction companies. This study predicted the business failure over the next three years using the 21 financial indicators based on the financial data of the past three years. The sample companies were classified into failed and normal companies using a finance-based definition of failure regardless of legal events. In order to handle the imbalance problem of the data, oversample technique based on SMOTE was used. Lastly, the six single classifier models' performances were compared by way of AUC values.

There are two contributions in this study. First, this study used a finance-based definition of failure in order to classify the sample companies regardless of legal events: bankruptcy, delisting, and default. The advantage of this definition is that it can be used to consider the financial risk companies may experience before legal events occur. In addition, a finance-based definition can prevent the problem of legal events occurring much later than the actual moment of failure. Second, this study considered the long-term business failure of construction companies in order to cover the relatively long duration of construction projects. At the early stages of construction projects, the long-term prediction model can predict whether construction companies will have financial risk until the end of the project or not. Therefore, the long-term prediction model can help project owners and other stakeholders to avoid the damage caused by business failure during the construction project.

This study used 21 financial ratios indicating the activity, leverage, liquidity, and profitability of construction companies. However, many studies [14,16,17,35] employed growth ratios to predict business failure in other companies. Therefore, in future study, we would include growth ratios as a financial indicator to improve the prediction performance. In addition, in the modeling method, a future study could extend the method to improve performance in long-term prediction.

References

[1] Horta I. M. and Camanho A. S. Company failure prediction in the construction industry. *Expert Systems with Applications*, 40:6253 – 6257, 2013.

[2] Tserng H. P., Lin G. F., Tsai L. K., and Chen P. C. An enforced support vector machine model for construction contractor default prediction. *Automation in Construction*, 20:1242 – 1249, 2011.

[3] National Information and Credit Evaluation Investors Service Rating performance 2015. Online:http://www.nicerating.com/disclosure/ratingPerFormance.do.

[4] Ng S. T., Wong J. M. W., and Zhang J. Applying Z-score model to distinguish insolvent construction companies in China. *Habitat International*, 35:599 – 607, 2011.

[5] Tserng H. P., Cheng P. C., Huang W. H., Lei M. C., and Tran Q. H. Prediction of default probability for construction firms using the logit model, *Journal of Civil Engineering and Management*, 20(2):247 – 255, 2014.

[6] Chen J. H. Developing SFNN models to predict financial distress of construction companies. *Expert Systems with Applications*, 39:823 – 827, 2012.

[7] Heo J. and Yang J. Y. AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing*, 24:494 – 499, 2014.

[8] Cheng M. Y., Hoang N. D., Limanto L., and Wu Y. W. A novel hybrid intelligent approach for contractor default status prediction. *Knowledge-Based Systems*, 71:314 – 321, 2014.

[9] Tserng H. P., Ngo T. L., Chen P. C., and Tran L. Q. A grey system theory-based default prediction model for construction firms. *Computer-Aided Civil and Infrastructure Engineering*, 30:120 – 134, 2015.

[10] Balcaen S. and Ooghe H. 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, 38:63 – 93, 2006.

[11] Tinco M. H. and Wilson N. Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, 30:394 – 419, 2013.

[12] Kisvalue, online: http://www.kisvlue.com.

[13] Pindado J., Rodrigues L., and Torre C. Estimating financial distress likelihood. *Journal of Business Research*, 61:995 – 1003, 2008.

[14] Li H. and Sun J. Ranking-order case-based reasoning for financial distress prediction. *Knowledge-Based Systems*, 21:868 – 878, 2008.

[15] Sun J. and Li H. Financial distress prediction based on serial combination of multiple classifiers. *Expert Systems with Applications*, 36:8659 – 8666, 2009.

[16] Hua Z., Wang Y., Xu X., Zhang B., and Liang L. Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications*, 33:434 – 440, 2007.

[17] Geng R., Bose I., and Chen X. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, 241:236 – 247, 2015.

[18] Ding Y., Song X., and Ze Y. Forecasting financial condition of Chinese listed companies based on support vector machine. *Expert Systems with Applications*, 34:3081 – 3089, 2008.

[19] Witten I. H. and Frank E. Data Mining: Practical Machine Learning Tools and Techniques, second ed. Morgan Kaufmann, San Francisco, 2005.

[20] Thammasiri D, Delen D., Meesad P., and Kasap N. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41:321 – 330, 2014.

[21] Zhou L. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems*, 41:16 – 25, 2013.

[22] Gao M., Hong X., Chen S., Harris C. J. A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. *Neurocomputing*, 74:3456 – 3466, 2011.

[23] Fernández A., Jesus M. J., and Herrera F. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. *International Journal of Approximate Reasoning*, 50:561 – 577, 2009.

[24] Son H. and Kim C. Early prediction of the performance of green building projects using pre-project planning variables: data mining approaches. *Journal of Cleaner Production*, 109:144 – 151, 2015.

[25] Chawla N. V., Bowyer K. W., Hall L. O., and Kegelmeyer W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321 – 357, 2002.

[26] Kim M. K., Kang D. K., and Kim H. B. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications*, 42:1074 – 1082, 2015

[27] Min J. H. and Lee Y. C. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, 28:603 – 614, 2005.

[28] Wu T. K., Huang S. C., and Meng Y. R. Evaluation of ANN and SVM classifiers as predictors to the diagnosis of students with learning disabilities. *Expert Systems with Applications*, 34:1846 – 1856, 2008.

[29] Barreto G. A. and Araújo A. F. R. Identification and Control of Dynamical Systems Using the Self-Organizing Map. *IEEE Transactions on Neural Networks*, 15(5):1244 – 1259, 2004.

[30] Sakthivel N. R., Sugumaran V., and Babudevasenapati S. Vibration based fault diagnosis of monoblock centrifugal pump using decision tree. *Expert Systems with Applications*, 37:4040 – 4049, 2010.

[31] Ravikumar S., Ramachandran K. I., and Sugumaran V. Machine learning approach for automated visual inspection of machine components. *Expert Systems with Applications*, 38:3260 – 3266, 2011.

[32] Balabin R. M., Safieva R. Z., and Lomakina E. I. Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. *Analytica Chimica Acta*, 671:27 – 35, 2010.

[33] Son H., Kim C., Hwang N., Kim C., and Kang Y. Classification of major construction materials in construction environments using ensemble classifiers. *Advanced Engineering Informatics*, 28:1 – 10, 2014.

[34] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceeding of the International Joint Conference on Artificial Intelligence*, pages 1137–1143, Montréal, Canada, 1995.

[35] Lin F., Liang D., Yeh C. C., and Huang J. C. Novel feature selection methods to financial distress prediction. *Expert Systems with Applications*, 41:2472 – 2483, 2014.