Expert Systems With Applications 83 (2017) 405–417

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Machine learning models and bankruptcy prediction

Flavio Barboza[a,*], Herbert Kimura[b], Edward Altman[c]

[a] Federal University of Uberlândia, Uberlândia, Minas Gerais 38408–100, Brazil
[b] Department of Management, Campus Darcy Ribeiro, University of Brasília, Brasília, Federal District, 70910–900, Brazil
[c] Leonard N. Stern School of Business, New York University, New York-NY, 10012-1126, USA

ARTICLE INFO

Article history:
Received 22 September 2015
Revised 4 March 2017
Accepted 3 April 2017
Available online 10 April 2017

JEL Classification:
C45
C52
C63
G33
L25

Keywords:
Bankruptcy prediction
Machine learning
Support vector machines
Boosting
Bagging
Random forest

ABSTRACT

There has been intensive research from academics and practitioners regarding models for predicting bankruptcy and default events, for credit risk management. Seminal academic research has evaluated bankruptcy using traditional statistics techniques (e.g. discriminant analysis and logistic regression) and early artificial intelligence models (e.g. artificial neural networks). In this study, we test machine learning models (support vector machines, bagging, boosting, and random forest) to predict bankruptcy one year prior to the event, and compare their performance with results from discriminant analysis, logistic regression, and neural networks. We use data from 1985 to 2013 on North American firms, integrating information from the Salomon Center database and Compustat, analysing more than 10,000 firm-year observations. The key insight of the study is a substantial improvement in prediction accuracy using machine learning techniques especially when, in addition to the original Altman's Z-score variables, we include six complementary financial indicators. Based on Carton and Hofer (2006), we use new variables, such as the operating margin, change in return-on-equity, change in price-to-book, and growth measures related to assets, sales, and number of employees, as predictive variables. Machine learning models show, on average, approximately 10% more accuracy in relation to traditional models. Comparing the best models, with all predictive variables, the machine learning technique related to random forest led to 87% accuracy, whereas logistic regression and linear discriminant analysis led to 69% and 50% accuracy, respectively, in the testing sample. We find that bagging, boosting, and random forest models outperform the others techniques, and that all prediction accuracy in the testing sample improves when the additional variables are included. Our research adds to the discussion of the continuing debate about superiority of computational methods over statistical techniques such as in Tsai, Hsu, and Yen (2014) and Yeh, Chi, and Lin (2014). In particular, for machine learning mechanisms, we do not find SVM to lead to higher accuracy rates than other models. This result contradicts outcomes from Danenas and Garsva (2015) and Cleofas-Sanchez, Garcia, Marques, and Senchez (2016), but corroborates, for instance, Wang, Ma, and Yang (2014), Liang, Lu, Tsai, and Shih (2016), and Cano et al. (2017). Our study supports the applicability of the expert systems by practitioners as in Heo and Yang (2014), Kim, Kang, and Kim (2015) and Xiao, Xiao, and Wang (2016).

1. Introduction

Financial institutions, fund managers, lenders, governments, and financial market players seek to develop models to efficiently assess the likelihood of counterparty default. Although default events behave stochastically, capital market information can be used to develop bankruptcy prediction models. For example, Altman (1968), in a seminal paper, applies multivariate statistical techniques, primarily discriminant analysis, to classify solvent and insolvent companies using financial statement data.

Credit risk arises due not only to bankruptcy events but also to the downgrading of the debt ratings of credit-related assets. Although default models have been studied for decades, the 2007/2008 financial crisis has made credit risk management a priority. However, Wang, Ma, and Yang (2014) suggest that there is no mature or definite theory of corporate failure. The lack of a theoretical framework within which to examine bankruptcy indicates the need for exploratory efforts to identify discriminant characteristics and predictive models for credit risk based on trial and error (Li & Sun, 2009; Wang et al., 2014; Zhou, Lai, & Yen, 2014).

* Corresponding author.
E-mail addresses: flmbarboza@ufu.br (F. Barboza), herbert.kimura@gmail.com (H. Kimura), ealtman@stern.nyu.edu (E. Altman).

http://dx.doi.org/10.1016/j.eswa.2017.04.006
0957-4174/

Researchers and practitioners have sought to improve bankruptcy forecasting models using various quantitative approaches. For example, Ohlson (1980) was one of the first researchers to apply logistic regression analysis to default estimation. In contrast to the model of Altman (1968), which generates a score by which to classify observations between good and bad payers, Ohlson's model (Ohlson, 1980) determines the default probability of the potential borrower.

Given the relative ease of running discriminant analysis and logistic regression, several subsequent studies have sought to perform similar tests (e.g. Hillegeist, Keating, Cram, and Lundstedt (2004), Upneja and Dalbor (2001), Griffin and Lemmon (2002), and Chen, Chollete, and Ray (2010)). However, Begley, Ming, and Watts (1996) argued that the popular models based on Altman (1968) and Ohlson (1980) had become inaccurate and suggested the need for enhancements in the modelling of default risk.

Academics and practitioners are exploring artificial intelligence and machine learning tools to assess credit risk amid advances in computer technology. Since credit risk analysis is similar to pattern-recognition problems, algorithms can be used to classify the creditworthiness of counterparties (Kruppa, Schwarz, Arminger, & Ziegler, 2013; Pal, Kupka, Aneja, & Militky, 2016), thus improving upon traditional models based on simpler multivariate statistical techniques such as discriminant analysis and logistic regression. Other methods have also been developed, offering new alternatives for credit risk analysis. Among these, we highlight machine learning methods. Support vector machines (SVMs)(Cortes & Vapnik, 1995), for example, generate functions similar to discriminant analysis, but they are not subject to series of assumptions and so are less restrictive. Other machine learning methods with wide applicability to predictive models have also been proposed, including default models such as boosting, bagging, and random forest models. Artificial neural networks (ANN) have been applied in many contexts as well. The incorporation of these machine learning algorithms seems promising. For example, Nanni and Lumini (2009) used Australian, German, and Japanese financial datasets to find that machine learning techniques, such as ensemble methods, lead to better classification than standalone methods.

Although many studies have analysed corporate solvency using modern computational techniques, Wang et al. (2014) found that the results did not identify the best method, since model performance depended on the specific characteristics of the classification problem and on the data structure (Duéñez Guzmán & Vose, 2013). Furthermore, Wang, Hao, Ma, and Jiang (2011) used ensemble methods (bagging, boosting, and stacking) coupled with base learners (logistic regression, decision trees, ANN, and SVM) to find that bagging outperformed boosting for all credit databases they analysed.

Several studies have dealt with the discussion of strengths and weaknesses of machine learning in many different disciplines, such as Subasi and Ismail Gursoy (2010) and de Menezes, Liska, Cirillo, and Vivanco (2017) in medicine; Laha, Ren, and Suganthan (2015); Maione et al. (2016) and Cano et al. (2017) in chemistry; Bernard, Chang, Popescu, and Graf (2017) in education; and Cleofas-Sánchez, García, Marqués, and Sénchez (2016); Heo and Yang (2014); Kim, Kang, and Kim (2015) and Gerlein, McGinnity, Belatreche, and Coleman (2016) in finance. However, our study does contribute to this debate.

First, our study focuses on the comparison of traditional statistical methods and machine learning techniques for predicting corporate bankruptcy. Although some papers have studied credit default and machine learning (Danenas & Garsva, 2015; du Jardin, 2016; Tsai, Hsu, & Yen, 2014; Wang et al., 2014; Zhou et al., 2014), new studies, exploring different models, contexts and datasets, are relevant, since results regarding the superiority of models are still inconclusive. The debate over the best models for predicting failure will probably continue in the short and medium terms, as new techniques are frequently being suggested and, particularly for the study of corporate bankruptcy, failure events are subject to myriad variables. In this context, for instance, with the advancement of technology, data scraping will allow the observation of new variables that could be relevant inputs to machine learning models and lead to different results.

Second, the variety of techniques and the applicability to practitioners can also be considered contributions of the study. By using raw data and considering standardized computer settings for the machine learning techniques, all our models can be easily replicated, not only by academics, but also by market practitioners. In this context, these models can be implemented in real world situations to address, for instance, the case of investors that could better understand and analyse strategic credit decisions, and the case of lender institutions that can improve their credit risk controls, based on results of machine learning models. Finally, we analyse a large database of corporate failure in the United States, by integrating data from 1985 and 2013 from the Salomon Center and Compustat. The use of a broad database of public companies, with more than 10,000 firm-year data records in the test set, is unusual in machine learning studies of corporate credit risk and can reveal relevant information of corporate bankruptcy in the North American environment. More specifically, various papers, such as Wang et al. (2014); Yeh, Chi, and Lin (2014); Zhao et al. (2014), and Xiao, Xiao, and Wang (2016), use a smaller number of observations of specific banks or credit card companies. Although results of these studies can convey information on adequacy of machine learning models, they are usually confined to specific characteristics of some financial institutions and their clients. In this context, results of our analysis can be more general, allowing for the understanding of default, not in a specific bank, but rather in the North American market for corporate loans. We highlight that, to the best of our knowledge, we did not find, in the machine learning literature, studies of corporate bankruptcy that investigate a similar number of observations, with all the techniques employed in our study.

Our work investigated the performance of different classification techniques by considering various machine learning algorithms applied to the practical problem of default prediction. In a comparative study, we used data from a training set of defaulted and non-defaulted firms covering 1985 to 2005 and a validation set covering 2006 to 2013, thus obtaining a confusion matrix. Overall accuracy indicators and area under the receiver operating characteristic (ROC) curve (AUC) were employed as performance metrics to compare the models. To evaluate the significance of the variables used in this study, its results were compared with those produced when the same models used only the Z-score variables. All the models showed lower accuracy when the number of variables was reduced, and the models with fewer variables produced higher type I and type II error rates.

The rest of the paper proceeds as follows. In Section 2, we briefly discuss the main machine learning models. In Section 3, we present the study's method and data. We discuss the classification results of the models in Section 4. In Section 5, we present final comments, discuss the implications of the study, including the strengths and weakenesses of the paper, and offer suggestions for future research.

2. Theoretical background

Machine learning methods are considered to be among the most important of the recent advances in applied mathematics, with significant implications for classification problems (Tian, Shi, & Liu, 2012). Machine learning techniques assess patterns in observations of the same classification and identify features that dif-