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-

ferentiate the observations of different groups. Machine learning studies are found across a wide range of research fields, including medicine (Noble, 2006; Subasi & Ismail Gursoy, 2010), engineering (Oskoei & Hu, 2008), and computing (Osuna, Freund, & Girosi, 1997). In this study, machine learning mechanisms are designed to distinguish between bankrupt and non-bankrupt companies based on firm characteristics such as profitability, liquidity, leverage, size, and growth measures. We compare applications of SVM, boosting, bagging, and random forest methods with artificial neural networks, logistic regression, and discriminant analysis. This section briefly reviews each of these mechanisms, considering each one's specific goals, mathematical modelling, and learning algorithms.

The solution of the credit analysis problem – specifically, of application scoring – involves an identification of the category (e.g. good vs. bad borrower, bankrupt vs. non-bankrupt firm) to which each observation belongs. The procedure is based on the definition of potential discriminant variables and the identification of weights or coefficients that can be used in mathematical functions that could segregate the groups.

2.1. Support vector machines

Following Noble (2006), the SVM optimisation model is based on the transformation of a mathematical function by another function, called the 'kernel', by which one identifies the greatest distance between the most similar observations that are oppositely classified.

A common criterion is whether the groups are completely separable, as this would allow the SVM to build a model with 100% accuracy. In finance, doing so is virtually impossible because economic variables are influenced by noise in empirical data and are often biased. For classification problems involving partially separable groups, the SVM method allows the inclusion of a margin of error (Zhou et al., 2014).

In general, the number of variables is not a constraint on the optimisation problem (Trustorff, Konrad, & Leker, 2010). The algorithm associated with the quantitative model establishes a classification mechanism, calibrating parameters using a training set (i.e. the algorithm learns from the training data). The resulting classification scheme can then be applied to predict the grouping or classification of new observations. The validation set is usually evaluated by comparing the classification given by the model with the actual group to which the observation belongs. The validation and training sets are independent: no observations are common between them (Yu, Yue, Wang, & Lai, 2010).

From Li, Wang, and He (2013), the optimisation problem can be summarised as

$$\text{Minimise} \quad \frac{1}{2}w^T w + C\sum_{i=1}^{M}\xi_i, \tag{1}$$

subject to

$$y_i[w^T\phi(x_i) + b] \geq 1 - \xi_i, \tag{2}$$

where $i = 1, 2, \ldots, M$, $\xi_i \geq 0$ are the margins of error related to classification cost C, y_i are the classifications in the training set, and $\phi(x)$ transforms space $\mathbb{R}^M$. One advantage of this technique is that $\phi(x)$ does not need to be known, since a kernel function ($K(x) = K(x_i, x_j)$) is applied so that $K(x) = \phi(x_i)^T.\phi(x_j)$.

The kernel function is predetermined in the algorithm and a solution to the optimisation problem (Eqs. (1) and (2)). The traditional kernel functions are

$$K(x_i, x_j) = < x_i, x_j >, \tag{3}$$

and

$$K(x_i, x_j) = e^{(-\gamma||x_i - x_j||^2)}. \tag{4}$$

where γ is a positive constant. Eq. (3) is called the linear kernel and (4) is the 'radial basis function' (RBF). The linear kernel does not provide strong predictability in non-separable datasets, due to the complexity of the empirical analysis, but the results are easily interpreted by users. Meanwhile, although the RBF kernel is difficult to analyse, or even discuss, it provides superior predictions in non-separable cases.

The SVM method is discussed in detail in Cortes and Vapnik (1995); Min and Lee (2005), and Yu et al. (2010).

2.2. Bagging

Bagging, also known as 'bootstrap aggregating', is a technique involving independent classifiers that uses portions of the data and then combines them through model averaging, providing the most efficient results concerning a collection (Breiman, 1996). Bagging creates random new subsets of data through sampling, with replacement, from a given dataset, generating confidence-interval estimates (Figini, Savona, & Vezzoli, 2016). The objective of bagging is to reduce the overfitting of a class within the model. Rather than using the collection to check if the model is overfitted, the training set is recombined to produce better classifiers.

Our bagging algorithm, based on Breiman (1996), follows these steps:

1. A random bootstrap set, t, is selected from the parent dataset.
2. Classifiers C_t are configured on the dataset from step 1.
3. Steps 1 and 2 are repeated for $t = 1, \ldots, T$.
4. Each classifier determines a vote,

$$C(x) = T^{-1}\sum_{t=1}^{T} C_t(x) \tag{5}$$

where x is the data of each element from the training set. In the last step, the class that receives the largest number of votes is chosen as the classifier for the dataset.

2.3. Boosting

The boosting technique consists of the repeated use of a base prediction rule or function on different sets of the initial set. Boosting builds on other classification schemes and assigns a weight to each training set, which is then incorporated into the model (Begley et al., 1996). The data are then reweighted. Boosting can apply the base classifier to find a model that better classifies the set, identified by a low error rate for the training set.

A derived algorithm, AdaBoost (adaptive boost) has proved successful for classification prediction (e.g. Kim & Upneja (2014)). AdaBoost initialises the weights of all m observations at $1/m$. Thus, the first sample is uniformly generated from the initial observations. After the training set, X_i, is extracted from X, a classifier Y_i is trained on X_i. The error rate is calculated, considering the number of observations of the training set. The new weight for each observation is based on the effectiveness of the classifier Y_i. If the error rate is greater than a random guess, the test set is discarded, and another set is generated with the original weights (initially $1/m$). If the error rate is satisfactory, the weights of the observation are updated according to the importance of the classifier. These new weights are then used to generate another sample from the initial observations. Our algorithm follows Heo and Yang (2014):

1. A distribution of weights, $w_1(i) = 1/m$ is created, where $i = 1, 2, \ldots, m$; and w_t is the iterative weighting ($t = 1, \ldots, T$),

$$w_{t+1}(i) = \frac{w_t(i)e^{\alpha_t(2I(y_i \neq h_t)-1)}}{w_t(i)e^{\alpha_t I(y_i \neq h_t)}}, \tag{6}$$

where $h_t = \text{argmax}|0.5 - \xi_t|$ is the error such that $\xi_t = \sum_{t=1}^{m} w^t(t) I(y_t \neq h_t(x_t))|$, and $I = 1$ when the measure was accurately computed, and 0 otherwise.
2. In each cycle, $\alpha_t = \frac{1}{2}\ln(\frac{1-\xi_t}{\xi_t})$ is recalculated.
3. The routine completes when $|0.5 - \xi_t| \neq \delta$, where δ is a predefined constant.
4. $Y(x)$ is evaluated for the completed boost

$$Y(x) = \text{sign} \sum_{t=1}^{T} \alpha_t h_t(x) \tag{7}$$

2.4. Random forest

The random forest technique (RF) is based on decision tree models, also known as generalised classification and regression trees' (CART). The model created by Breiman (2001) has a level of precision similar to that of AdaBoost and, depending on the set, can provide better results than boosting can (Kruppa et al., 2013). It is particularly robust and allows for the presence of outliers and noise in the training set (Yeh et al., 2014). Finally, RF identifies the importance of each variable in the classification results. Therefore, it provides not only the classification of observations, but also information about the determinants of separation among groups Maione et al. (2016). The RF technique follows an approach similar to bagging, as it repeatedly generates classification functions based on subsets. However, RF randomly selects a subset of characteristics from each node of the tree, avoiding correlation in the bootstrapped sets (Booth, Gerding, & McGroarty, 2014; Cano et al., 2017; Yeh et al., 2014). The forest is built for several subsets that generate the same number of classification trees. The preferred class is defined by a majority of votes, thus providing more precise forecasts and, most importantly, avoiding data overfitting (Breiman (2001)).

Our RF algorithm follows Yeh et al. (2014):

1. Create random subsets of the parent set, composed of an arbitrary number of observations and different features.
2. Each subset from step 1 produces a decision tree, and all elements of the set have a label (correct or not).
3. For each element, the forest takes a large number of votes. The class with the most votes is chosen as the preferred classification of the element.

A more detailed discussion on random forests, including a more rigorous mathematical description, is found in Booth et al. (2014); Breiman (2001), and Calderoni, Ferrara, Franco, and Maio (2015).

Table 1 shows a summary of some papers that investigate applications of machine learning techniques in financial issues, mainly in bankruptcy prediction. We follow similar research approach.

2.5. Artificial neural networks

Artificial neural networks (ANN) are among the most popular artificial intelligence techniques and have inspired other computational classification models. This method establishes an analogy with human neural processing (Park, Kim, & Lee, 2014; Tsai et al., 2014). Many non-linear relationships can be analysed using ANN methods. More recently, it has been shown that machine learning methods tend to provide better classification results (Freund and Schapire, 1997; Kim and Kang, 2010; and Kruppa et al., 2013). For instance, Zhao et al. (2014) used German credit data to build a credit-scoring model using ANN. The authors argued that ANN with back propagation predicts credit scores accurately. They found 87% efficiency from the classification and concluded that their

Table 1
Summary of relevant studies in the paper context. ACC means the accuracy of the model presented; Obs. Amount is the dataset size; Attrib shows how many explanatory variables were applied.

Paper	Dataset	Techniques	Acc	Obs. amount	Attrib.	Prediction concern	Benefits	Weakness
Tsai (2014)	Taiwan	NN, SVM, Boosting, Bagging	86%	440	95	Bankruptcy	Combined techniques	(1) No comparison with statistical method; (2) Many attributes with no feature selection
Booth et al. (2014)	DAX Stocks	Linear Reg., Reg. Tree, NN, SVR, RF	0.01 (MAPE)	13 years	30	Price trending	(1) Feature selection; (2) Practical use of the models is the authors' concern; (3) Moving Averages are involved.	(1) Neither the attributes selected nor the number of attributes selected are indicated. (2) Unique method (RF).
Danenas (2015)	USA (EDGAR)	SVM, NN, LR	93%	9 years	51	Risk of bankruptcy	(1) SVM outperforms; (2) Feature selection	(1) No change/growth variables are used; (2) SVM is less stable than others.
Cleofas (2016)	UCI, Iran, Poland, Spain, Thomas, UCSD, USA	NN, SVM, LR	78%	240–8200	12–30	Fin. distress	(1) Different NN architectures are tested.	(1) No comments about features; (2) Financial Distress definition are missing.
Heo(2014)	Korea	AdaBoost, NN, SVM, DT	94%	30,000	12	Bankruptcy	(1) AdaBoost outperforms; (2) Variable analysis is shown; (3) Modelling by company size.	(1) Comparisons with Z-score is simplified. (2) One industry is studied. (3) No change/growth variables.
Kim(2014)	Korea	SMOTE, SVM, Adaboost, GMBoost	95%	10,000	30	Bankruptcy	(1) Overview about imbalanced data issue; (2) Feature selection (2) AUC as ACC Measure	(1) Bank failures only on dataset; (2) No change/growth variables. (3) No comparisons with traditional models.
Liang (2016)	Taiwan	SVM, KNN, CART, NN, NB	82%	480	190	Bankruptcy	(1) Corporate governance and growth measures are included the model; (2) AUC is one of the performance measures; (3) Discuss the importance of new variables	(1) Selected features not mentioned; (2) Only SVM is tested.
Xiao (2016)	Germany, Australia	SVM, Bagging, DT, LR	–	1000;700	20; 14	Credit scoring	(1) Models are tested for different base classifiers; (2) Applies in a case study.	No change/growth variables.
Wang (2014)	Poland; Other	Boosting, Bagging, DT, NN, NB, SVM, LR	82%; 87%	240; 132	30; 24	Bankruptcy	(1) Feature selection; (3) Designed research is clear; (3) Improvements for boosting technique.	(1) No change/growth variables.(2) Variables analysis is missing. (3) Old dataset cannot be useful.

Table 2
Amount of failure (F) and Non-failure (NF) companies by year and per industry: (1) Agriculture,... ,(10) Wholesale trade.

Industry	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)		(9)		(10)		**Total**	
Year	F	NF	F	NF	F	NF	F	NF	F	NF	F	NF	F	NF	F	NF	F	NF	F	NF	F	NF
1985		10		354	1	48	6	373		302		229	1	346		50		372	1	56	9	2140
1986		12	1	323		52	3	352		282		227	1	370	2	47		340		56	7	2061
1987		13		313		56	1	328	1	271		244	1	358	1	43	1	314		59	5	1999
1988		14		291	1	46	1	309		236		232		337		36	1	291	1	58	4	1850
1989		14		292		44	2	290	1	233		213	2	310		34	1	264		64	6	1758
1990		14	1	284	1	44	6	266		226		205	2	299	3	32	4	255	1	68	18	1693
1991		16		278		43	1	255		226		205	3	315	1	29	2	234		59	7	1660
1992		14		261	1	45	6	240	1	221		209	3	355	2	26	3	228		60	16	1659
1993		16		247		48	4	219	1	219		234	2	363		24	1	232		60	8	1662
1994		16		246		51	2	246	1	208		254	7	342		23	1	210		74	11	1670
1995		16	2	242		55	2	241	1	203		274	2	342	3	26		213		90	10	1702
1996		14		241		60	5	263	1	195		289	3	320		27	1	231		85	10	1725
1997		16		211	1	66	6	258	1	189		281	3	289	1	19	3	218	3	95	18	1642
1998		17		191	1	68	12	276	2	180		264	8	263	5	21	1	218		92	29	1590
1999		14	1	184		54	10	226	4	167	1	269	9	249	10	19	5	206	3	99	43	1487
2000	1	13		184	4	52	27	222		153		239	11	232	12	20	14	187	4	106	73	1408
2001		15		179	6	52	28	224	2	152	1	226	9	218	14	21	15	200	4	102	79	1389
2002		17		165	1	47	22	209	1	147		209	5	205	5	23	5	197	3	91	42	1310
2003		13		170	1	43	16	205	1	143		191	5	196	3	22	4	167	1	96	31	1246
2004		14		164		43	11	160	1	142		193	1	182		21	2	153	1	83	16	1155
2005		13		172	1	40	5	182	1	138		181	2	167		17	6	132		77	15	1119
2006		12		165		42		173		139		183	3	159	1	17	1	122	2	75	7	1087
2007		10		155	1	44	6	164		123	1	179	2	145	2	17	2	119		81	14	1037
2008		10	1	158	1	45	24	155	1	124		164	6	138	4	16	4	109	3	77	44	996
2009		13	1	144		46	3	122	2	119		164	1	135	6	14	2	98	2	34	17	889
2010		15		150		46	5	125	1	115		159	1	130	2	14	1	96		20	10	870
2011		16	1	136	1	46	5	119		115		163		122	4	11	1	87		26	12	841
2012		18		142	2	45	6	113		105		153	1	121	5	10	3	97	2	30	19	834
2013		17	1	136		45	1	105	1	101		140		118	1	7	2	78		19	6	766
Total	1	422	9	6,509	24	1,463	226	6,778	25	5,377	3	6,389	94	7,443	87	742	86	6,013	31	2,040	586	41,155

model was somewhat better than traditional prediction mechanisms.

The ANN algorithm we use is similar to that used in Wang et al. (2011) and Zhao et al. (2014). The model is a structure (network) created in layers with linkages among nodes (neurons). Input variables determine the first layer of the modelling system, and the final layer provides the output (dependent) variable. Here, the dependent variable is the classification of 'bankrupt' (one year before filing date) and 'non-bankrupt' companies. Since default probability is also important in the model, we use a real number between 0 (bankrupt) and 1 (non-bankrupt).

Three problems with the ANN technique have been identified by Zhao et al. (2014): (i) its performance for unbalanced data is poor because it tends to classify more observations in classes with more data and, reducing the test set's forecasting performance; (ii) model accuracy improves as the training set becomes larger, but the validation is insufficient to provide a satisfactory error rate; and (iii) selecting the hidden layers is difficult, given the relationship between computing time (i.e. more time is required for more layers) and higher predictability. We adopt a scheme similar to that in Zhao et al. (2014) to address these issues and apply ANN.

2.6. Discriminant analysis and logistic regression

Multivariate discriminant analysis (MDA) was a breakthrough in credit risk assessment that occurred when Altman (1968) presented a study of bankruptcy among manufacturing firms that achieved relevant classification results. The method is based on the minimisation of the variance among observations of the same group and the maximization of the distance between observations of different groups (Mahmoudi & Duman, 2015). The method produces a score, and an observation is classified into a group depending on the score relative to an arbitrary cut-off value. The restrictive assumptions of MDA, such as requiring normally distributed variables and sensitivity to outliers, made the logistic regression (LR) a more popular alternative as a multivariate model for application scoring and, subsequently, for credit risk modelling (du Jardin, 2016). Not only are the assumptions less restrictive, but LR also produces a result in the [0, 1] interval that can be interpreted as the probability of a given observation being a member of a specific group (de Menezes et al., 2017). Several studies (e.g. Kruppa et al. 2013; Trustorff et al. 2010) have shown that these traditional multivariate methods are not as efficient or accurate as are more recent machine learning techniques for credit risk classification.

3. Data and method

We collected financial data on American and Canadian companies covering 1985 to 2013 using Compustat. Information on firm insolvency was collected from NYU's Salomon Center database. A subset covering 1985 to 2005 was extracted to provide the training set, which included information on 449 companies that filed for bankruptcy during this period as well as information on the same number of non-bankruptcy firms. Insolvent firms in the training set include all companies in the database that filed for bankruptcy during this period and for which financial data were available three years prior to filing. The solvent firms were randomly chosen and were limited to companies that did not file for bankruptcy during the entire period (1985–2005) and for which financial data for at least two consecutive years were available. We selected the same number of solvent and insolvent firms, following Altman (1968), who also considered a balanced set. Table 2 show the number of solvent and insolvent firms in each year and industry.

We chose the predictive variables based on two important studies: the seminal paper by Altman (1968) and a review of organisational performance by Carton and Hofer (2006). Five variables fol-

Table 3
Predictive variables of the default classification model for one year prior to bankruptcy. If bankruptcy filing occurred less than one semester after fiscal year end, data were collected from the previous fiscal year. Some measures require changes over time to compute these variables; we use data up to three years prior to bankruptcy.

Variable	Formula
X1	$\frac{\text{Net working capital}}{\text{Total assets}}$
X2	$\frac{\text{Retained earnings}}{\text{Total assets}}$
X3	$\frac{\text{Earnings before interest and taxes}}{\text{Total assets}}$
X4	$\frac{\text{Market value of share} * \text{number of shares}}{\text{Total debt}}$
X5	$\frac{\text{Sales}}{\text{Total assets}}$
OM	$\frac{\text{Earnings before interest and taxes}}{\text{Sales}}$
GA	$\frac{\text{Total assets}_t - \text{Total assets}_{t-1}}{\text{Total assets}_{t-1}}$
GS	$\frac{\text{Sales}_t - \text{Sales}_{t-1}}{\text{Sales}_{t-1}}$
GE	$\frac{\text{Number of employees}_t - \text{Number of employees}_{t-1}}{\text{Number of employees}_{t-1}}$
CROE	$\text{ROE}_t - \text{ROE}_{t-1}$ where $ROE = \frac{\text{Net income}}{\text{Common Stockholders' equity}}$
CPB	$\textit{Price-to-Book}_t - \textit{Price-to-Book}_{t-1}$ where $P/B = \frac{\text{Market value per share}}{\text{Book value per share}}$

low the relevant financial dimensions in Altman (1968): liquidity (X1), profitability (X2), productivity (X3), leverage (X4), and asset turnover (X5). To evaluate the potential impact of other dimensions in predicting bankruptcy, we also included indicators with a greater influence on financial performance models in the short term: growth of assets (GA), growth in sales (GS), growth in the number of employees (GE), operational margin (OM), change in return on equity (CROE), and change in price-to-book ratio (CPB) (Carton & Hofer, 2006). Data were rearranged as variables (see Table 3).

The validation set contains a randomly chosen group of 133 bankrupt firms and 13,300 companies considered solvent from 2006 (which it is not included in the training set) and 2013. For bankrupt companies, we included all those firms with data available in the database at least one year before filing. If the event occurred during the first half of the fiscal year, the data were collected from the second preceding fiscal year. In another group, all 13,167 solvent (non-bankrupt) companies were selected from a random year within this period for the test set.

All variables were included in the models at their original values. No transformation, such as normalization, was conducted. Although this procedure may reduce the predictive power of the models, we aimed to analyse the adequacy of machine learning techniques without relying on specific or special treatment of data in the sample. The use of originally available data without any transformation was also followed, for example, by Cleofas-Sánchez et al. (2016); Heo and Yang (2014); Tsai et al. (2014). For a brief visualization of the descriptive statistics, Table 4 presents a summary of the full sample. We also analysed the potential impact of missing values and found no relevant difference in the data, as depicted in the Table 4.

Eight techniques were applied:

- Bagging,
- Boosting,
- Random forest (RF),
- SVM with two kernels: linear (SVM-Lin) and radial basis function (SVM-RBF),
- Artificial neural networks (ANN),
- Logistic regression (Logit), and
- MDA.

We implemented the models using the R statistical software packages. Specifically, this study used ada, e1071, mboost, randomForest, MASS, aod, and nnet to implement bagging, SVM, boosting, random forest, MDA, Logit, and ANN, respectively. For the machine learning models, the regression trees were used as base learners for bagging and random forest, while recursive partitioning trees were applied by SVM. It is important to note that these learners are presented in the default package settings.

It is important to analyse, especially for the use of traditional statistical techniques, namely, logistic regression, potential correlations among variables. Table 5 depicts correlations in different scenarios, since firm-year data from non-defaulting companies are chosen randomly.

Almost all correlations are not relevant, except those between X1 (liquidity) and X2 (profitability), X2 and X3 (productivity) and, to a lesser degree, X1 and X3, in the training sample with eleven variables. The other samples do not show a relevant high correlation between variables. We investigated the database and found that the high correlation observed in the specific training sample derives from an outlier related to a defaulted firm. Since the number of bankruptcies is relatively small, and the study using 11 variables reduces the number of observations with non-missing data, the correlation was sensitive to the outlier in this particular sample. We chose to maintain the correlated variables in the models, since they were also used in the seminal paper from Altman (1968) and the results of predictions using different models can be compared. It is important to highlight that the new metrics proposed in this work did not correlate with any other variable, suggesting that they would be potential candidates to contribute to a better prediction of bankruptcy. To check for robustness, we conducted a study of the correlation matrix, excluding the outlier, and a study excluding one correlated variable (X2). The prediction results do not substantially change. The results are presented in Tables 8–10, respectively, of the Supplementary Material.

The ROC curve was calculated for all models for the training and validation sets by using the ROCR package, providing a critical analysis of the evolution of machine learning. The AUC also provided a criterion of accuracy for the validation set: the AUC had to be more than 0.5 for the model to be acceptable, and the closer it was to 1, the stronger its predictive power.

Table 4
Descriptive statistics (Minimum, 1st. quarter, median, mean, 3rd quarter, maximum, and standard deviation (SD) of the full sample. First, data including missing values (NA's). Second, without NA. Third, only data from bankruptcy firms. Fourth, only from non-bankrupt firms.

Variable	X1	X2	X3	X4	X5	GA	GS	GE	OM	CR	CPB	BK
Min	−15415	−134863	−23957.5	0.00	−17.195	−1	−58.66	−1	−30175.7	−166842.9	−107120.79	−1
1st Qu	0.031	−0.38	−0.04	0.73	0.462	−0.04	−0.02	−0.06	−0.017	−0.06	−0.59	1
Median	0.212	0.09	0.064	1.83	1.006	0.07	0.1	0.02	0.06	0	−0.03	1
Mean	−1.071	−15.24	−0.731	15.27	1.29	2.11	1.19	0.25	−3.554	−0.32	−0.41	0.9959
3rd Qu	0.41	0.3	0.122	5.28	1.596	0.23	0.26	0.14	0.126	0.04	0.43	1
Max	16.238	140.58	1625	188244	13203	97584	15054	2699	394.474	39864.33	107500.13	1
SD	60.19829	510.1601	53.04549	NA	25.87659	NA	NA	NA	NA	NA	NA	0.0903936
NA's	0	0	0	60753	0	28766	40979	77422	14754	84603	87795	0
Variable	X1	X2	X3	X4	X5	GA	GS	GE	OM	CR	CPB	BK
Min	−3800.375	−24638	−23957.5	0.00	−11.5385	−0.9995	−50.286	−1	−30175.7	−166842.9	−107120.79	−1
1st Qu	0.062	−0.198	−0.006	0.68	0.6257	−0.0377	−0.023	−0.0563	−0.005	−0.06	−0.55	1
Median	0.238	0.144	0.072	1.6	1.1	0.0665	0.093	0.0217	0.061	0	−0.03	1
Mean	−0.036	−2.858	−0.215	6.85	1.2895	0.2375	0.856	0.186	−2.65	−0.53	−0.62	0.9941
3rd Qu	0.418	0.344	0.126	4.17	1.6258	0.1991	0.24	0.1406	0.122	0.04	0.4	1
Max	16.238	140.582	35.917	188244	434.9835	2405	15054	2699	394.474	26773.44	18555.57	1
SD	18.38568	129.8665	54.37543	428.4017	2.299579	8.833552	55.07403	7.770179	107.917	395.095	273.8783	0.1088771
Failures	X1	X2	X3	X4	X5	GA	GS	GE	OM	CR	CPB	BK
Min	−4.057	−33.5285	−2.8514	0.0000	0.0000	−0.8786	−0.9832	−0.912	−20.2059	−538.1912	−878.61	−1
1st Qu	−0.07863	−0.52582	−0.06963	0.0694	0.6171	−0.1645	−0.08415	−0.14348	−0.07995	−1.9442	−1.2222	−1
Median	0.0725	−0.1574	0.0004	0.1874	1.1184	−0.0375	0.00765	−0.0414	0.0004	−0.3272	−0.2987	−1
Mean	−0.02921	−0.60389	−0.05185	0.406	1.2459	0.1238	0.42396	0.77861	−0.27476	−3.9153	−2.8037	−1
3rd Qu	0.1873	0.01365	0.04097	0.4324	1.611	0.1251	0.24672	0.08282	0.03692	−0.0128	0.3538	−1
Max	0.779	1.9423	0.4904	6.5718	7.8175	13.4138	48.1326	311.5	0.3839	363.9792	69.6994	−1
SD	0.440936	2.201727	0.2224851	0.6662472	0.9397091	0.8645904	2.923344	13.30826	1.464162	39.35201	38.36061	0
Non−Failures	X1	X2	X3	X4	X5	GA	GS	GE	OM	CR	CPB	BK
Min	−3800.375	−24638.000	−23957.500	0.00	−11.5385	−0.9995	−50.286	−1.0000	−30175.700	−166842.90	−107120.79	1
1st Qu	0.063	−0.199	−0.006	0.68	0.6284	−0.0376	−0.023	−0.0563	−0.005	−0.06	−0.55	1
Median	0.240	0.1450	0.072	1.62	1.1019	0.0665	0.093	0.0220	0.061	0.00	−0.03	1
Mean	−0.037	−2.874	−0.216	6.89	1.2916	0.2383	0.860	0.1847	−2.666	−0.52	−0.62	1
3rd Qu	0.418	0.346	0.126	4.20	1.6272	0.1996	0.240	0.1413	0.121	0.04	0.41	1
Max	16.238	140.582	35.917	188244.00	434.9835	2405.0000	15054.000	2699.0000	394.474	26773.44	18555.57	1
SD	18.44065	130.2547	54.53805	429.6827	2.305391	8.859838	55.2385	7.759349	108.239527	396.270799	274.689357	0

Table 5
Correlation matrices for main datasets: full sample, training and testing samples using Altman's 5 variables, and training and testing samples for all 11 variables.

Painel A: all data available											
All data	X1	X2	X3	X4	X5	GA	GS	GE	OM	CR	CPB
X1											
X2	0.72***										
X3	0.09***	0.47***									
X4	0	0	0								
X5	−0.19***	−0.29***	−0.25***	0							
GA	0	0	0	0.01***	−0.01*						
GS	0	0	0	0	0	0.07***					
GE	0	0	0	0	0	0.03***	0.02***				
OM	0.01***	0.01***	0.01*	0	0.01***	−0.01**	0	0			
CR	0	0	0	0	0	0	0	0	0		
CPB	0	0	0	0	0	0	0	0	0	0	
BK	0	0	0	0	0	0	0	0	0	0	0
Painel B: five variable models											
Train	X1	X2	X3	X4	X5	**Test**	X1	X2	X3	X4	X5
X1						X1					
X2	0.68***					X2	0.59***				
X3	0.19***	0.38***				X3	0.24***	0.35***			
X4	0	0	0			X4	0	0	0		
X5	−0.26***	−0.32***	−0.26***	0		X5	−0.07***	−0.15***	−0.22***	0	
BK	0	0	0	0	0	BK	0	0	0	0	0
Painel C: eleven variable models											
Train	X1	X2	X3	X4	X5	GA	GS	GE	OM	CR	CPB
X1											
X2	0.71***										
X3	0.12***	0.71***									
X4	0	0	0								
X5	−0.31***	−0.45***	−0.28***	−0.02***							
GA	0	0	0	0.01**	0						
GS	0	0	0	0	0.01*	0.07***					
GE	0	0	0	0	0	0.07***	0.05***				
OM	0.01***	0.01***	0.01**	−0.02***	0.01***	−0.01**	0	0			
CR	0	0	0	0	0	0	0	0	0		
CPB	0	0	0	0	0	0	0	0	0	0	
BK	0	0	0	0.01*	0	0	0	−0.01**	0	0	0
Test	X1	X2	X3	X4	X5	GA	GE	GS	OM	CR	CPB
X1											
X2	0.58***										
X3	0.30***	0.26***									
X4	0.01	0	-0.05***								
X5	-0.36***	-0.31***	0.01	-0.06***							
GA	0	0	0	0	0.01						
GE	0	0	0	0	-0.01	0.05***					
GS	0	0	0	0	0	0.07***	0.01				
OM	0.13***	0.09***	0.29***	-0.05***	0.04***	0	0	0			
CR	0	0	0	0	0	0	0	0	0		
CPB	0	-0.01	-0.03**	0.01	-0.02*	0	0	0.01	-0.01	0	
BK	0	-0.01	0	0.02**	-0.01	0	0	0	-0.01	0	0

Note: * $p < 0.05$; ** $p < 0.01$; and *** $p < 0.001$

Two commonly applied performance rates (Kim & Upneja, 2014; Wang, Ma, Huang, & Xu, 2012) were calculated: the true positive rate (TPR) or sensitivity and the true negative rate (TNR) or specificity, which are equivalent to 1−type I error and 1−type II error, respectively. The predictive power or accuracy (ACC) was calculated as the number of accurate classifications divided by the total number of elements in the validation set. These indicators are equivalent to those proposed by Altman (1968)); hence, we can compare them with his outcomes directly. The variables are given by:

$$\text{Sensitivity} = \text{TPR} = 1 - \text{Type I Error} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{Specificity} = \text{TNR} = 1 - \text{Type II Error} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

where TP is True Positive, that is, bankrupt firms classified correctly and TN is True Negative, that is, non-bankrupt firms classified correctly. Sensitivity has values close to 1 when type I error is low, and specificity is close to 1 when type II error is low. For bankruptcy, there is a preference for higher sensitivity because this translates into losses for lenders, whereas specificity is the threshold for gain. Fig. 1 illustrates our methodology.

4. Results

Table 6 shows the outcomes for traditional and machine learning models in the training and the test sets. We used a standard MacBook Air (4GB DDR3L RAM Memory, 64GB of flash storage, 1.7GHz Intel Core i5 processor, and Mac OS X as operating system) with the R software, version 3.1.1 installed, and all packages cited below.

The bagging and RF techniques show high accuracy in the training phase. This outcome was expected, since both use decision trees, which can cause model overfitting in the training set. However, this does not mean that they are good models, as

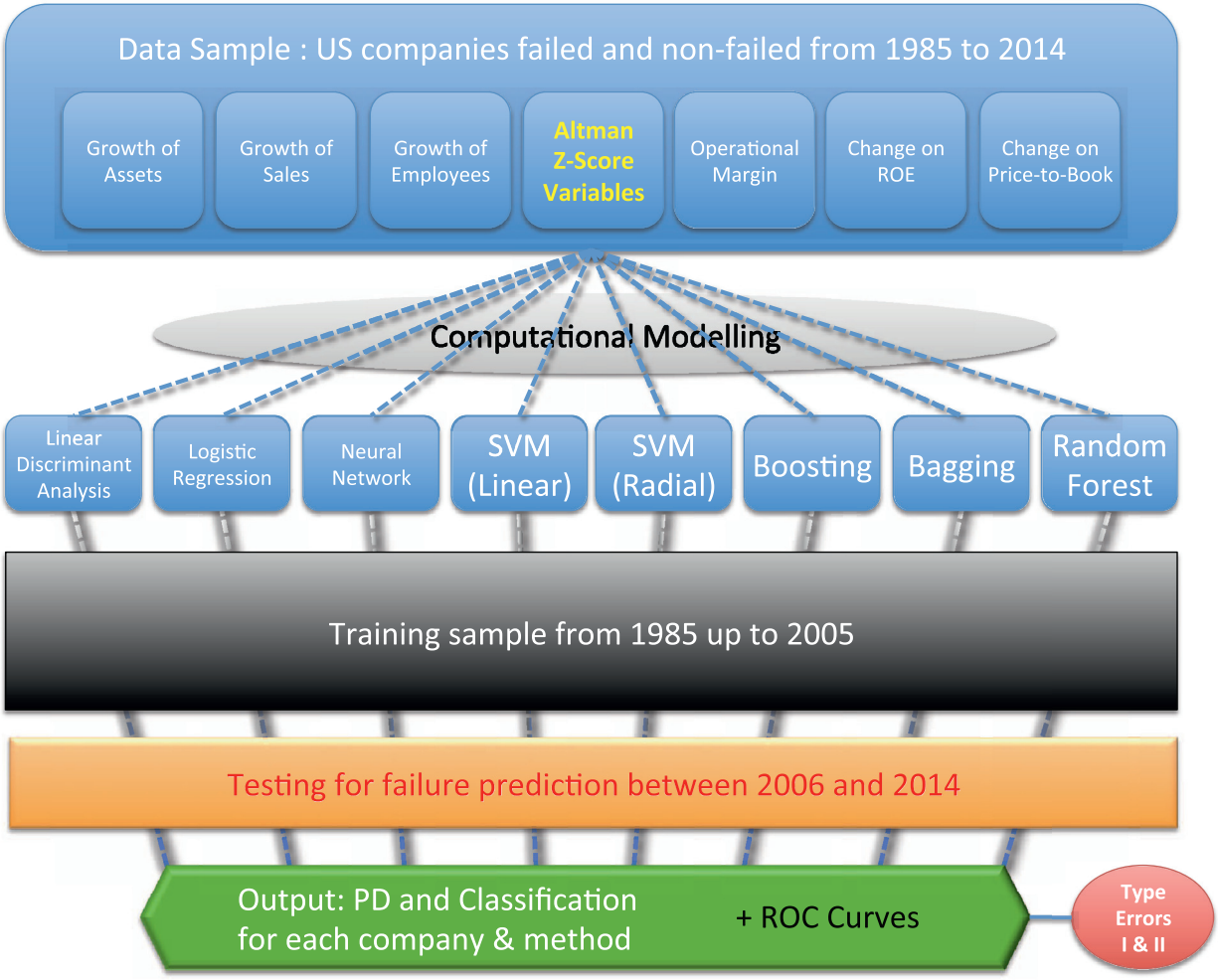

Fig. 1. Graphic abstract: 11 variables were selected, including Z-score variables and growth measures, change variables, and margins suggested by Carton and Hofer (2006). Machine learning models used: a support vector machine using linear and radial basis function kernels, boosting, bagging, and random forest. Traditional models used: artificial neural networks, logistic regression, and multivariate discriminant analysis. The results are presented as confusion matrices and ROC curves.

Table 6
The 13,300 firms tested in our models. Type I error means the portion of bankrupt companies that were predicted to be non-bankrupt. Type II error means the portion of non-bankrupt companies that were predicted to be bankrupt. AUC is the area under the ROC curve, and ACC is total estimated accuracy.

Training sample Model	TP	TN	FP	FN	Type I Error (%)	Type II Error (%)	AUC (%)	ACC (%)
SVM-Linear	419	306	143	30	6.68	31.85	NA	80.73
SVM-RBF	421	376	73	28	6.24	16.26	NA	88.75
Boosting	434	430	19	15	3.34	4.23	NA	96.21
Bagging	448	447	2	1	0.22	0.45	NA	99.67
Random forest	449	449	0	0	0.00	0.00	NA	100.00
Neural networks	431	331	118	18	4.01	26.28	NA	84.86
Logit	414	329	120	35	7.80	26.73	NA	82.74
MDA	361	221	228	88	19.60	50.78	NA	64.81
Testing sample Model	TP	TN	FP	FN	Type I Error (%)	Type II Error (%)	AUC (%)	ACC (%)
SVM-Linear	123	9,389	3,778	10	7.52	28.69	67.2	71.52
SVM-RBF	105	10,505	2,662	28	21.05	20.22	85.17	79.77
Boosting	108	11,417	1,750	25	18.80	13.29	92.97	86.65
Bagging	110	11,284	1,883	23	17.29	14.30	92.48	85.67
Random forest	111	11,468	1,699	22	16.54	12.90	92.92	87.06
Neural networks	124	9,582	3,585	9	6.77	27.23	90.08	72.98
Logit	118	10,028	3,139	15	11.28	23.84	90.10	76.29
MDA	86	6,854	6,313	47	35.34	47.95	63.68	52.18

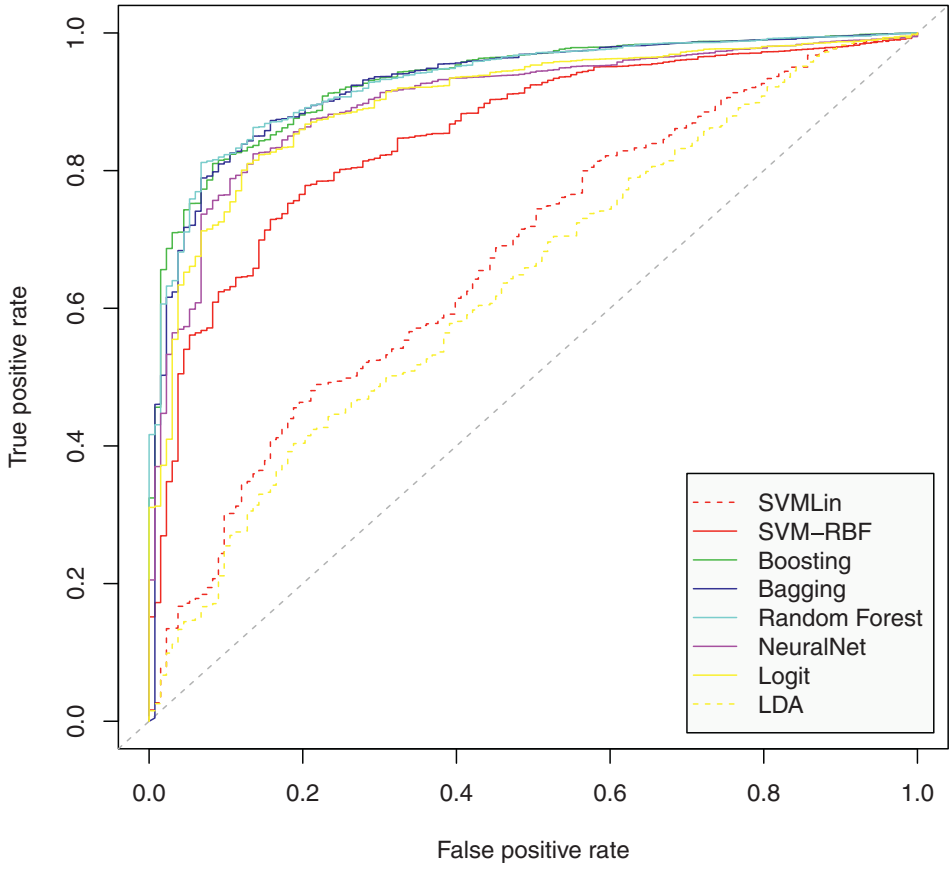

Fig. 2. ROC curve of all models to predict bankruptcy from 2006 up to 2014, with data from the Compustat database and NYU Salomon center. MDA and SVM-Lin show less and boosting, bagging and RF greater predictive power.

evidenced by the significantly poorer outcomes of the validation set. Table 6 shows the metrics (AUC, ACC, Type I and Type II errors) discussed above when the models were applied to the test set. Machine learning models outperform traditional models measured on the validation set, except SVM-Lin. The linear structure applied to separate the two classes (bankrupt vs. non-bankrupt) causes this weak performance, since MDA, which uses a linear process, displays a similar accuracy problem. Thus, non-linear techniques are necessary when more variables are used to predict bankruptcy. While the ANN model had a lower type I error (6.8%), followed by SVM-Lin (7.5%), their type II error is considerably higher than that of other machine learning models (i.e. ANN [27.2%], SVM-Lin [28,7%] as opposed to boosting [13.3%], bagging [14.3%], and RF [12.9%]). Boosting provides the most accurate AUC rate (92.9%), but RF returns the lowest type II error (12.9%) and the best total accuracy rate (87.1%). The machine learning model's errors and prediction rates are all better than those of the MDA model. While Logit performs acceptably in classifying bankruptcy firms (11.3% type II error), misclassification of solvent companies indicates a poor model (23.8%).

Bagging shows reasonable performance and may provide an interesting alternative to the more computationally intense machine learning systems. Type II error was the second-best rate among the eight models tested (1.4% above RF) and third-best for AUC and ACC. Fig. 2 shows the ROC curve of each model. The machine learning models show significant superiority over MDA and Logit, except SVM-Lin, which is most similar to MDA, due to their inherently linear models, as discussed above. While it is difficult to confirm a single preferred technique from the curves, bagging, boosting, and RF are the most promising candidates.

We performed several tests with the selection of variables. Using only the original Z-score (Altman, 1968) variables leads to significantly inferior performance in terms of predictive power in this more recent database. The success rate was comparable to that for the model that used only the variables selected by Carton and Hofer (2006). Our outcomes show the importance using more explanatory variables, since bankruptcy may be reflected by many different indicators. Table 7 shows the results using only the variables from Altman (1968). In this case, 14 companies were added to the test set to replace companies that lacked sufficient data to calculate growth rates; therefore, data from only one year before the event were used. The 14,553 healthy companies were resampled randomly, following the proposed methodology.

The results show reduced performance among all forecasting models. While ANN and SVM-RBF were slightly improved for AUC (by 1% and 4%, respectively), ANN decreased in type I error. The SVM-Lin model, although presenting reduced accuracy, improved its prediction of bankrupt companies, which also occurred with the MDA model. This is most likely due to the reduction in variables allowing a more linear cluster and thus a better performance for these linear models. Bagging achieved the highest AUC rate. However, the type I errors of the three most precise models (boosting, bagging and RF) were significantly impacted (approximately 23%), whereas MDA (19%) and Logit (9.5%) outperformed these models for this measure. Type II error shows the opposite result with a large separation ($\approx$ 20% for Logit and $\approx$ 36% for MDA), generating a significantly better ACC. Following the methodology outlined here, we also investigated different periods: training in pre-crisis periods and validation in crisis periods (and vice versa); training for five consecutive years and tests one year later; and training for three

Table 7
Predictions of the models applied using the variables of the popular Altman model (Altman, 1968). All methods show loss of predictability, sensibility, and specificity. This shows the importance of including new variables in the models.

Model	TP	TN	FP	FN	Type I Error (%)	Type II Error (%)	AUC (%)	ACC (%)
SVM - Linear	141	9,203	5,350	6	4.08	36.76	66.31	63.56
SVM - RBF	129	10,900	3,653	18	12.24	25.01	89.54	75.03
Boosting	113	12,575	1,978	34	23.13	13.59	91.25	86.31
Bagging	116	12,452	2,101	31	21.09	14.44	91.49	85.5
Random forest	112	12,536	2,017	35	23.81	13.86	91.15	86.04
Neural networks	138	10,047	4,506	9	6.12	30.96	91.09	69.29
Logit	133	9,659	4,894	14	9.52	33.63	86.18	66.61
MDA	119	7,254	7,299	28	19.05	50.15	67.4	50.16

consecutive years and validation in the three subsequent years. Detailed results are omitted for brevity. The various outcomes were inferior to those of the model described here, perhaps indicating that historical data are not relevant to current events. However, in all tests without exception, the three machine learning techniques – boosting, bagging and RF – showed better outcomes in all tests, with the latter usually being the best.

5. Conclusions

Bankruptcy prediction is associated with credit risk, which has been thrust into the spotlight due to the recent financial crisis. Machine learning models have been very successful in finance applications, and many studies examine their use in bankruptcy prediction. The Altman and Ohlson models are still relevant, due not only to their predictive power but also to their simple, practical, and consistent frameworks. Few studies can improve on their results concerning forecasting accuracy or the simplicity of the models.

Regarding accuracy ratios, our results show that the traditional models (MDA, LR, and ANN) have lower predictive capacity (between 52% and 77%) than the machine learning models (71% to 87%), corroborating Wang et al. (2012), Breiman (2001), Kim and Upneja (2014) and Chen et al. (2010). New studies can adapt these machine learning techniques for other credit risk studies, such as general default events not limited to bankruptcy.

However, machine models are not perfect. The SVM-Lin results show that it is difficult to address non-separable datasets, which produces more misclassifications. SVM-RBF, a non-linear kernel, reduced error rates but had weaker performance than other machine learning models. Moreover, SVM models took significantly more computational processing time than others did; however, this was not an important issue because no model took more than one minute to run. Although bagging, boosting, and RF incorporate similar procedures, RF generally produced better accuracy and error rates. Output variability is a critical problem typically found in ANNs, whereas the machine learning models produced stable solutions.

We now highlight some strengths of the study. First, we argue that significant predictive accuracy was achieved by using machine learning models under restrictive conditions such as existence of high correlated variables, outliers, and missing values. Since we use raw data, without making any transformation or adjustments to the variables, results suggest that machine learning techniques can easily be applied and generate substantial classification accuracy, when compared to traditional mechanisms such as linear discriminant analysis, logistic regression, and artificial neural networks.

Another differential element of the study is the use of metrics that reflect growth or change in some variables, which are not usually incorporated into predictive models of bankruptcy. Following arguments from Carton and Hofer (2006), we argue that failure of firms is likely to follow from difficulties over time and not just in the year prior to bankruptcy. Instead of using the time series or survival analysis approaches, our procedure easily incorporates information about short-term evolution of a variable in a machine learning model. Results show that, by adding variables that depict a dynamic yet simple behaviour of firms, machine learning techniques may improve predictive accuracy.

We also highlight that our using of an extensive database of defaults is not common in papers related to corporate bankruptcy. Many studies that investigate default are restricted to a specific credit product of a specific financial institution. Since we use broad training and testing samples, the results can be useful in analysing bankruptcy of the overall North American corporate credit market. The number of observations also reduces problems of overfitting. Accuracy rates close to 90% in the testing sample represents evidence of the suitability of machine learning models to analyse corporate default. It is also important to discuss the weaknesses of the study. As we used default parameters of the algorithms implemented in R packages, we do not take advantage of the full capacity of the models. Nonetheless, even by using simple settings of the algorithms, results show a relevant superiority of machine learning techniques.

In addition, our study did not focus on feature selection, which is a common procedure in recent studies that explore a high number of variables. However, as usual in studies regarding bankruptcy, one has access to a very limited number of variables. In particular, even though we use a large number of observations, the databases supplied a limited number of variables. Therefore, the impact of feature selection would not be prominent in our study. Pal et al. (2016) argues that feature selection in the finance context "depends upon the individual judgment of the analyst or group decision-making. This makes the theoretical basis for the feature selection limited and less reliable". Finally, another limitation of the study is not considering different classification costs, similarly to Cleofas-Sánchez et al. (2016); Liang, Lu, Tsai, and Shih (2016), and Mahmoudi and Duman (2015). We find that, especially for prediction of bankruptcy, accuracy should not be the only performance metric, and future research should focus on adjusting classification models by considering different impacts of type I and type II errors.

Credit risk applications – specifically, default prediction – should be further investigated, particularly in efforts to obtain models related to macroeconomic variables. Several papers have found relationships between default and macroeconomic variables (e.g. Ali & Daly, 2010; Bonfim, 2009; Chen & Wu, 2014; Yurdakul, 2014). We chose not to use these macroeconomic data as inputs in the models because the effects of our metrics in terms of firm-specific measures produced relevant outcomes. This choice may constitute a limitation that could be explored in a subsequent study. However, given the scope of this paper, further research is needed to incorporate the impact of macroeconomic variables such as sustainability, governance, sovereign risk, credit spreads, and firm performance. Another limitation of this study is its validation procedure. The overfitting analysis of machine learning models is not well explained in the literature. Questions about which tools and theories are most appropriate for such analyses remain open to further investigation.

We can highlight major implications of the study, from two different perspectives. First, for academics, we tested machine learning models using an unusually large sample for the study of corporate bankruptcy. We use a very representative sample, including several sectors, with more than 13,000 observations. Thus, considering the large sample and the high predictive accuracy, results are not restricted to a specific bank or credit portfolio and can be useful to analyse failure in North American corporate loans. Second, for practitioners, the results of the study add to a growing literature (e.g. Danenas & Garsva, 2015; Kruppa et al., 2013; López Iturriaga & Sanz, 2015) related to better performance of machine learning techniques when compared to traditional approaches that are widespread in the credit industry, such as logistic regression. These results should encourage decision makers to test and consider the use of machine learning models in their databases. Although practitioners could be concerned about explanatory reasons to validate their model, the complexity of the bankruptcy phenomenon would suggest that machine learning could be an important tool to aid credit risk analysis. If the goal of the decision maker is to predict and not necessarily explain (Efron & Hastie, 2016), then the use of estimates of prediction error should be the focus and relative contribution of predictors would not be a matter of concern. In this context, results show that machine learning could be a powerful ally to make decisions about corporate loans.

Future studies should extend the analysis to incorporate the growth rates and/or time effects of all variables, including the growth measures themselves, to evaluate the impact of time on default events. The outcomes should be applied to individual financial institutions, while considering specific institutional aspects such as ratings, credit losses, economic capital, and credit spreads.

For practitioners, this study's outcomes are interesting in that they reveal how using computational learning techniques can enhance the predictive power of credit risk models. Banks and risk managers can investigate these machine learning models, which could improve their credit risk analysis and thus help them achieve better profitability with lower credit risk exposure.

Acknowledgements

This document is a collaborative effort. We thank Santander Bank, CAPES Foundation – Ministry of Education of Brazil (Process number 1766/2014-07), Centro Estadual de Educação Tecnológica Paula Souza (CEETEPS) and CNPq (Process numbers 409725/2013-7 and 310666/2016-3) for their financial support. The authors also thank Credit & Debt Markets Research Program from Salomon Center, New York University and Brenda Kuehne (Credit & Debt Markets Research Specialist) for the list of bankrupt firms.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.eswa.2017.04.006

References

Ali, A., & Daly, K. (2010). Macroeconomic determinants of credit risk: Recent evidence from a cross country study. *International Review of Financial Analysis, 19*(3), 165–171.
Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609.
Begley, J., Ming, J., & Watts, S. (1996). Bankruptcy classification errors in the 1980s: An empirical analysis of Altman's and Ohlson's models. *Review of Accounting Studies, 1*(4), 267–284.
Bernard, J., Chang, T.-W., Popescu, E., & Graf, S. (2017). Learning style identifier: Improving the precision of learning style identification through computational intelligence algorithms. *Expert Systems with Applications, 75*, 94–108.
Bonfim, D. (2009). Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *Journal of Banking & Finance, 33*(2), 281–299.
Booth, A., Gerding, E., & McGroarty, F. (2014). Automated trading with performance weighted random forests and seasonality. *Expert Systems with Applications, 41*(8), 3651–3661.
Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Calderoni, L., Ferrara, M., Franco, A., & Maio, D. (2015). Indoor localization in a hospital environment using random forest classifiers. *Expert Systems with Applications, 42*(1), 125–134.
Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., Thapa, A., & Barr, A. (2017). Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Systems with Applications, 72*, 151–159.
Carton, R., & Hofer, C. (2006). *Measuring organizational performance*. Edward Elgar Publishing.
Chen, J., Chollete, L., & Ray, R. (2010). Financial distress and idiosyncratic volatility: An empirical investigation. *Journal of Financial Markets, 13*(2), 249–267.
Chen, P., & Wu, C. (2014). Default prediction with dynamic sectoral and macroeconomic frailties. *Journal of Banking & Finance, 40*, 211–226.
Cleofas-Sánchez, L., García, V., Marqués, A., & Sénchez, J. (2016). Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing, 44*, 144–152.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
Danenas, P., & Garsva, G. (2015). Selection of support vector machines based classifiers for credit risk domain. *Expert Systems with Applications, 42*(6), 3194–3204.
Efron, B., & Hastie, T. (2016). Computer Age Statistical Inference.
Figini, S., Savona, R., & Vezzoli, M. (2016). Corporate default prediction model averaging: A Normative linear pooling approach. *Intelligent Systems in Accounting, Finance and Management, 23*(1–2), 6–20.
Freund, Y., & Schapire, R. E. (1997). A decision-Theoretic generalization of on-Line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.
Gerlein, E. A., McGinnity, M., Belatreche, A., & Coleman, S. (2016). Evaluating machine learning classification for financial trading: An empirical approach. *Expert Systems with Applications, 54*, 193–207.
Griffin, J. M., & Lemmon, M. L. (2002). Book-to-Market equity, distress risk, and stock returns. *The Journal of Finance, 57*(5), 2317–2336.
Duéñez Guzmán, E. A., & Vose, M. D. (2013). No free lunch and benchmarks. *Evolutionary Computation, 21*(2), 293–312.
Heo, J., & Yang, J. Y. (2014). AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing, 24*, 494–499.
Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies, 9*(1), 5–34.
du Jardin, P. (2016). A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research, 254*(1), 236–252.
Kim, M.-J., & Kang, D.-K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications, 37*(4), 3373–3379.
Kim, M.-J., Kang, D.-K., & Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications, 42*(3), 1074–1082.
Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and adaboosted decision tree models. *Economic Modelling, 36*, 354–362.
Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications, 40*(13), 5125–5131.
Laha, D., Ren, Y., & Suganthan, P. (2015). Modeling of steelmaking process with effective machine learning techniques. *Expert Systems with Applications, 42*(10), 4687–4696.
Li, H., & Sun, J. (2009). Gaussian case-based reasoning for business failure prediction with empirical data in china. *Information Sciences, 179*(1–2), 89–108.
Li, S., Wang, M., & He, J. (2013). Prediction of banking systemic risk based on support vector machine. *Mathematical Problems in Engineering, 2013*, 1–5.
Liang, D., Lu, C.-C., Tsai, C.-F., & Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research, 252*(2), 561–572.
López Iturriaga, F. J., & Sanz, I. P. (2015). Bankruptcy visualization and prediction using neural networks: A study of u.s. commercial banks. *Expert Systems with Applications, 42*(6), 2857–2869.
Mahmoudi, N., & Duman, E. (2015). Detecting credit card fraud by modified fisher discriminant analysis. *Expert Systems with Applications, 42*(5), 2510–2516.
Maione, C., de Paula, E. S., Gallimberti, M., Batista, B. L., Campiglia, A. D., Jr, F. B., & Barbosa, R. M. (2016). Comparative study of data mining techniques for the authentication of organic grape juice based on ICP-MS analysis. *Expert Systems with Applications, 49*, 60–73.
de Menezes, F. S., Liska, G. R., Cirillo, M. A., & Vivanco, M. J. (2017). Data classification with binary response through the boosting algorithm and logistic regression. *Expert Systems with Applications, 69*, 62–73.
Min, J., & Lee, Y. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications, 28*(4), 603–614.
Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 36*(2), 3028–3033.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565–1567.
Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109–131.
Oskoei, M., & Hu, H. (2008). Support vector machine-Based classification scheme for myoelectric control applied to upper limb. *IEEE Transactions on Biomedical Engineering, 55*(8), 1956–1965.
Osuna, E., Freund, R., & Girosi, F. (1997). Training support vector machines: An application to face detection. In *Computer vision and pattern recognition, 1997. proceedings., 1997 ieee computer society conference on* (pp. 130–136). IEEE.
Pal, R., Kupka, K., Aneja, A. P., & Militky, J. (2016). Business health characterization: A hybrid regression and support vector machine analysis. *Expert Systems with Applications, 49*, 48–59.
Park, H., Kim, N., & Lee, J. (2014). Parametric models and non-parametric machine learning models for predicting option prices: Empirical comparison study over KOSPI 200 index options. *Expert Systems with Applications, 41*(11), 5227–5237.
Subasi, A., & Ismail Gursoy, M. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications, 37*(12), 8659–8666.
Tian, Y., Shi, Y., & Liu, X. (2012). Recent advances on support vector machines research. *Technological and Economic Development of Economy, 18*(1), 5–33.
Trustorff, J.-H., Konrad, P. M., & Leker, J. (2010). Credit risk prediction using support vector machines. *Rev Quant Finan Acc, 36*(4), 565–581.
Tsai, C.-F., Hsu, Y.-F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing, 24*, 977–984.
Upneja, A., & Dalbor, M. C. (2001). An examination of capital structure in the restaurant industry. *International Journal of Contemporary Hospitality Management, 13*(2), 54–59.
Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications, 38*(1), 223–230.
Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems, 26*, 61–68.
Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications, 41*(5), 2353–2361.
Xiao, H., Xiao, Z., & Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. *Applied Soft Computing, 43*, 73–86.
Yeh, C.-C., Chi, D.-J., & Lin, Y.-R. (2014). Going-concern prediction using hybrid random forests and rough set approach. *Information Sciences, 254*, 98–110.
Yu, L., Yue, W., Wang, S., & Lai, K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications, 37*(2), 1351–1360.
Yurdakul, F. (2014). Macroeconomic modelling of credit risk for banks. *Procedia - Social and Behavioral Sciences, 109*, 784–793.
Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2014). Investigation and improvement of multi-layer perception neural networks for credit scoring. *Expert Systems with Applications, in press*.
Zhou, L., Lai, K. K., & Yen, J. (2014). Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation. *International Journal of Systems Science, 45*(3), 241–253.