Journal of Computational Optimization in Economics and Finance ISSN: 1941-3971
Volume 3, Number 3 © 2011 Nova Science Publishers, Inc.

A COMPARISON OF ALTERNATIVE METHODOLOGIES FOR CREDIT RISK EVALUATION

Dimitrios Niklis[1], Michael Doumpos[1*] and Chrysovalantis Gaganis[2]

[1] Technical University of Crete, Department of Production Engineering and Management,
Financial Engineering Laboratory, Greece
[2] University of Crete, Department of Economics, Greece

ABSTRACT

Credit risk refers to the likelihood that a firm or individual borrower will fail to meet a debt obligation. Credit risk evaluation is a very challenging and important problem in the domain of financial risk management. There are different methods and approaches for constructing credit risk assessment rating systems. The aim of this paper is to perform an empirical comparison of different popular techniques using a data set of Greek companies from the commercial sector. For this purpose three different methodologies are used, namely logistic regression, support vector machines, and the UTADIS (UTilités Additives DIScriminantes) multicriteria method. The results show that even with a considerable imbalanced data set with a small number of defaults, all methods provide good results. The UTADIS multicriteria method outperforms the two other techniques. Ensemble models are also tested, but are found to provide only marginal improvements.

Keywords: Credit risk evaluation, Multicriteria techniques, Logistic regression, Support vector machines, Ensemble models

1. INTRODUCTION

Credit risk assessment is a very challenging and important topic in the domain of financial risk management. The field of credit risk modeling has developed rapidly over the past decades to become a key component in the risk management systems at financial institutions (Lopez and Saidenberg, 2000). Credit risk can be defined as the potential risk that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms (BIS, 2004). There are many factors that have increased the need for accurate credit risk measurement. Among others, these include: (i) a worldwide structural increase in the number of defaults, (ii) a trend towards disintermediation by the highest quality and largest

[*] Author for correspondence: E-mail: mdoumpos@dpem.tuc.gr

borrowers, (iii) more competitive margins on loans, (iv) a declining value of real assets (and thus collateral) in many markets, and (v) a dramatic growth of off-balance sheet instruments with inherent default risk exposure (McKinsey, 1993), including credit risk derivatives (Altman and Saunders, 1998). Accounting-based creditscoring systems have been the most popular tools for credit risk assessment and rating.

Such systems combine key accounting variables, which are properly weighted to produce either a credit risk score or a probability of default. If the credit risk score, or probability, attains a value above a critical benchmark, a loan applicant is either rejected or subjected to increased scrutiny (Altman and Saunders, 1998). Altman et al. (1977) developed the commonly used ZETA discriminant model. Platt and Platt (1991) used logistic regression to test whether industry relative accounting ratios, are better predictors of corporate bankruptcy. Smith and Lawrence (1995) used a logit model to find the variables that offer the best prediction of a loan moving into a default state. Despite their good performance, multivariate accounting-based credit scoring systems developed using statistical methods have been subject to criticism. They are based on book value accounting data models which have the disadvantage to be static and thus often fail to follow the changes in the economic and business environment. Furthermore, most of these models are linear, thus failing to model the available information accurately, as linearity does not always exist among explanatory variables. Due to these limitations, non-parametric methods have become popular over the past couple of decades. Among others, one can mention techniques such as neural networks (Altman et al., 1994; Piramuthu, 1999; Atiya, 2001; Baesens et al., 2003; Westet al., 2005; Angelini et al.., 2008), rough sets (Dimitras et al., 1999), support vector machines (Huang et. al, 2004; Stecking and Schebesch, 2003), multicriteria decision aid (Bugera et al., 2002; Doumpos et al., 2002; Kou et al., 2005; Hu, 2009; Doumpos and Zopounidis, 2011), and data envelopment analysis (Troutt et al., 1996; Cielen et al., 2004).

The aim of the study is the evaluation of different techniques and modeling settings in developing credit scoring models. In particular, three popular techniques are taken into consideration, namely logistic regression, support vector machines, and the UTADIS multicriteria method.

The analysis is based on a sample of Greek commercial firms, spanning the period 2006–2009. We examine the stability and robustness of the results using a bootstrap sampling approach, focusing on different settings with regard to the size of the model fitting (training) samples and their composition (number of default vs non-default observations). The construction of ensemble models is also tested and their results are compared against the individual models developed on the basis of the full sample.

The rest of article is organized as below. Section 2 outlines the basic characteristics and features of the different methodologies used in the analysis. Section 3 discusses the data, while section 4 presents the empirical results. Finally, Section 5 concludes the article, summarizes the main findings of this research and proposes some future research directions.

2. Classification Methods

The development of credit risk assessment and rating systems is based on classification methods. The objective is to fit a model that discriminates default from non-default cases as accurately as possible. In this study three popular methods are used, including logistic regression, support vector machines, and the UTADIS multicriteria method. Logistic regression is the most commonly used approach in the domain of credit risk modeling. It leads to a linear classification model which is easy to construct and understand. Support vector machines (SVMs) have become a popular statistical learning approach with numerous applications in classification, regression, and clustering problems. SVMs provide a common theoretical basis for the development of both linear and non-linear model. Finally, the UTADIS multicriteria method uses linear programming to construct additive models, which are monotone with respect to the input variables. The following subsections provide a brief outline of the selected techniques.

2.1. Logistic Regression

Logistic regression (LR) is the most widely used statistical approach for building credit scoring and rating models. In LR, the probability of non-default for a firm *i* is modeled based on a set of independent variables through the logistic function:

$$P(\mathbf{x}_i)=\frac{1}{1+e^{-(\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_m x_{im})}}$$

with $\mathbf{x}_i=(x_{i1},\ldots,x_{im})$ representing the input data vector for firm *i* on a set of *m* independent variables, α is the constant term and $\beta_1,\ldots,\beta_m$ are the regression coefficients of the independent variables. On the basis of the posterior probability estimates, each company is classified as default or non-default using an optimal probability cut-off point, which is specified so that the type I and type II errors[1] are minimized.

2.2. Support Vector Machines

Support vector machines (SVMs) have become an increasingly popular non-parametric methodology for developing classification models. In a dichotomous classification setting, and assuming a linear classification model, the objective is to construct a linear decision function $f(\mathbf{x}_i)=\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_m x_{im}$ that distinguishes the two classes as best as possible. The decision function classifies observation *i* in the class of positive cases (i.e., the non-default group) if and only if $f(\mathbf{x}_i)>0$. The analysis of the generalization performance of the decision function has shown that the optimal model *f* is the one that maximizes the margin induced in the separation of the classes, which is defined in terms of the vector of coefficients $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_m)$ as follows $2/\|\boldsymbol{\beta}\|$ (Vapnik, 1998). Therefore, given a training sample of *n* observations, the maximization of the margin can be performed through the solution of the following quadratic programming problem:

[1] Type I error refers to the classification of a default firm as a non-default one. On the other hand, type II error refers to the classification of a non-default firm as a default one.

$$\begin{aligned}
&\text{minimize} && \tfrac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta} + C\mathbf{e}^\top\mathbf{s} \\
&\text{subject to:} && \mathbf{Y}(\mathbf{X}\boldsymbol{\beta} + \alpha) + \mathbf{s} \geq \mathbf{e} \\
& && \mathbf{s} \geq \mathbf{0},\ \boldsymbol{\beta}, \alpha \in \mathbb{R}
\end{aligned}$$

where $\mathbf{Y}$ is a $n \times n$ diagonal matrix with the class labels in its diagonal (1 for the non-defaulted cases in -1 for the defaulted ones), $\mathbf{X}$ is a $n \times m$ matrix with the training data, $\mathbf{e}$ is a vector of ones, $\mathbf{s}$ is a vector of non-negative slack variables associated with the misclassification of the training objects when the classes are not linearly separable, and $C > 0$ is a parameter used to penalize the classification errors.

To generalize a linear SVM model to a non-linear one, the problem data are mapped to a higher dimensional space *H* (feature space) through a transformation of the form $\mathbf{x}_i\mathbf{x}_j^\top \to \phi(\mathbf{x}_i)\phi^\top(\mathbf{x}_j)$. The mapping function ϕ is implicitly defined through a symmetric kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)\phi^\top(\mathbf{x}_j)$. Popular choices for the kernel function include the polynomial kernel, the radial basis function (RBF) kernel, the sigmoid kernel, etc. (Schölkopf and Smola, 2002). The representation of the data using the kernel function enables the development of a linear model in the feature space *H*. Since *H* is a non-linear mapping of the original data, the developed model is non-linear in the original input space. The model is developed by applying the above linear analysis to the feature space *H*.

In this study we explore the development of both linear and non-linear SVM models with an RBF kernel. The width of the RBF kernel was selected through a cross-validation analysis to ensure the proper specification of this parameter. A similar analysis was also used to specify the trade-off constant *C*. All the data used during model development were normalized to zero mean and unit variance.

2.3. The UTADIS Multicriteria Method

The UTADIS method leads to the development of an additive value function that is used to score the firms and decide upon their classification. The developed additive value function has the following general form:

$$V(\mathbf{x}_i) = \sum_{j=1}^{m} w_j u_j(x_{ij}) \in [0,1]$$

where w_j is a non-negative trade-off constant for criterion *j* and $u_j(x_j)$ is the corresponding marginal value function normalized between 0 and 1. The marginal value functions provide a mechanism for decomposing the aggregate result (global value) in terms of individual assessments to the criteria level.

The value function provides an aggregate score $V(\mathbf{x}_i)$ for each firm *i*. To classify firms in their original groups (default or non-default), it is necessary to estimate a value threshold *t*, such that firms with global value at least equal to *t* are classified in the non-default group and all others are assigned in the default class.

The estimation of the additive value function and the cut-off threshold is performed through linear programming techniques. The objective of the method is to develop the additive model that minimizes the classification errors for the firms in the training sample. Detailed description of the mathematical programming formulation used in the UTADIS method can be found in the works of Zopounidis and Doumpos (1999) and Doumpos and Zopounidis (2002).

3. Data and Variables

3.1. Training and Testing Samples

The data were obtained from the financial database of ICAP[2]. The sampleconsists of 10,468 non-defaulted firm-year observations and 248 defaulted cases. All the firms in the sample belong to the commercial sector. The time period of the analysis covers the years 2006–2009. There are two groups of companies (defaulted and non-defaulted). Table 1 presents the observations per year and category.

Table 1. Sample observations by year and category

Year	Non-defaulted	Defaulted	Total
2006	2748	52	2800
2007	2846	53	2899
2008	2731	99	2830
2009	2143	44	2187
Total	10468	248	10716

The available data were partitioned into two disjoint data sets, namely a training and a testing sample. The model is fitted on the training data and its classification performance is analyzed on the basis of the testing sample. Generally, one can adopt different approaches to implement a training-testing scheme for assessing the performance of credit rating models. Barnes (1990) points out that due to inflationary effects, technological changes and numerous other reasons (e.g., changing accounting policies), it is unreasonable to expect that the distributional cross-sectional parameters of the financial ratios will remain stable over time. Thus, a realistic validation approach would require the evaluation of the model in future period, since this approach more closely reflects a real world setting. As Espahbodi and Espahbodi (2003) mention that "*After all, the real test of a classification model and its practical usefulness is its ability to classify objects correctly in the future. While cross-validation and bootstrapping techniques reduce the over-fitting bias, they do not indicate the usefulness of a model in the future*."

Therefore, in this study, in order to consider the case of population drifting (i.e. change of population over time) and determine whether the models remain stable over different time periods, we split the sample in to two distinct datasets. The first consists of data from the

[2] ICAP is the largest company in the Business Information and Consulting Services sector in Greece

period 2006–2007 and serves as a training sample. The second contains data from the subsequent two years (i.e. 2008 and 2009) and serves as a testing sample. Overall, the training data consist of 5,699 observations including 5,594 from the non-default group and 105 from the default class. The testing sample includes 4,874 non-defaulted cases and 143 defaulted ones (overall 5,017 firm-year observations). It should be noted that, due to the crisis that started to emerge in Greece during 2008, the default rate is higher in the testing sample compared to the training sample (2.9% in the testing sample vs 1.8% in the training sample).

3.2. Variable Selection

The use of financial ratios in order to evaluate business failures and also credit risk is very common. One of the first researchers who used financial ratios in order to predict the bankruptcy of companies was Beaver (1966). A drawback of his analysis was the use of each ratio per time and the creation of a cut off point for each one of them. Altman (1968) used a combination of six different ratios in order to classify the firms between healthy and distressed. Ohlson (1980) examined nine different ratios using logistic regression. Piramuthu (1999) used 18 different combinations of variables in order to classify a set of firms into those that would default and those that would not default. Atiya (2001) used 5 ratios in order to predict corporate bankruptcies.

The choice of the appropriate financial ratios is a challenging issue. First, there is a plethora of ratios that can be used as proxies for the same financial attributes (i.e., profitability, solvency, liquidity, etc.) and is often unclear which the best selections that can be made are. Second, using a large number of ratios increases the time and the cost of data collection and management. Third, a large set of ratios can lead to multicollinearity problems (Gaganis et al., 2007).

Hamer (1983) pointed out that the selected set of variables should be constructed on the basis of: (a) minimizing the cost of data selection, and (b) maximizing the applicability of the model. However, it is not easy to determine how many ratios a particular model should consider.

If a very small set of attributes is used, the model will not include all the relevant information, whereas the use of too many attributes could lead to two problems: (a) overfitting the training sample, and (b) time and effort spent to insertdata (Kocagil et al., 2002).

Some studies start from a large list of potentially useful attributes, which are later reduced through a statistical selection process such as hypotheses testing or multivariate data analysis (Emel et al., 2003). However, Palepu (1986) criticizes such an approach and argues that "*this method of variable selection is arbitrary and leads to the statistical overfitting of the model to the sample at hand*". Therefore, to avoid such criticisms, and at the same time enhance the applicability of the model, we include ratios that cover all aspects of business cycle of a company together with ones that their contribution is found statistically significant in order to classify firms according to previous studies.

Overall, the analysis is based on a set of 11 indicators (7 financial and 4 non-financial ratios), which are shown in Table 2.

Table 2. Variable descriptions

Panel A: Financial Ratios			
Category	Variables	Relationship to default	Previous Studies
Management efficiency	Short-term liabilities*365 / Cost of Sales (STL/CS)	+	Lacher et al. (1995); Lee et al. (1996)
	Accounts receivable*365 / Sales (AR/S)	+	McKee (2003); Hamerle et al. (2006); Altman and Narayanan (1997)
	Inventories / Cost of sales (I/CS)	+	Ahn et al. (2000); Cielen et al. (2004); Karels and Prakash (1987)
Profitability	Profit before tax / Total assets (PBT/TA)	–	Bryant (1997); McKee (2003); Min and Lee (2005); Bonfim (2009)
	Financial expenses / Sales (FE/S)	+	Lee et al. (1996); Park and Han (2002); Shin and Lee (2002).
Solvency	Quick assets / Short-term liabilities (QA/STL)	–	Piramuthu et al. (1998); Dimitras et al. (1999); Cielen et al. (2004); Ko L.J, et al.(2007)
	Total liabilities/Total assets (TL/TA)	+	Kolari et al. (2002); Swicegood and Clark (2001); Tung et al. (2004); Greco et al. (1998)
Panel B: Non-financial indicators			
	Logarithm of employees (LOGE)	–	Leshno and Spector (1996); Tung et al. (2004)
	Exports indicator (EXP)	–	Becchetti and Sierra (2003); Yurdakul and Tansel (2004)
	Imports indicator (IMP)	–	
	Representations indicator (REPR)	–	

We have four categories of financial ratios (efficiency, profitability, liquidity and financial leverage) and four non-financial indicators that are crucial for commercial companies. In the next paragraph we will briefly discuss the relevance of the selected attributes in the context of credit risk evaluation.

Management efficiency ratios are typically used to analyze how well a company uses its assets and liabilities. Efficiency ratios are important because an improvement in the ratios usually translates to improved profitability. We have selected three ratios in this category, which are positive related to credit risk, in the sense that the higher their value, the higher is the probability of default.

The profitability indicators are used to assess ability of a firm to generate earnings as compared to its expenses and other relevant costs incurred during a specific period of time. The profitability ratios considered in this study include the return on assets ratio and ratio of financial expenses to sales. The first one is negative related to credit risk in contrast with the second one which is positively associated to the probability of default.

Finally, the category of solvency indicators includes two ratios related to the liquidity and the financial leverage of the firms. Liquidity determines a company's ability to pay off its short-term debt obligations. In this study the quick ratio (Current assets-Inventories/Short-term liabilities) is used which is negative related to credit risk. On the other hand, financial leverage provides an indication of the long-term solvency of a firm. Here the ratio of total liabilities to total assets is used which is positive related to credit risk.

Apart from financial indicators there should be a consideration of other factors that affect the operation of a firm. Here two factors are examined:

1. Logarithm of employees. This is an indicator of the size of a company, which has been shown in past studies to be negatively associated to the probability of default.
2. Activity indicator. For commercial companies it is important to take into consideration the type of their activities. In this study, in accordance with ICAP's modeling approach, the activities of the companies in the sample as characterized as exporting, importing, or as representative (i.e., companies that are local resellers of products of foreign companies). For each of these three classes of business activities, three binary indicators are used to describe the sample observations.

3.3. Univariate Analysis

Table 3 presents some descriptive statistics for the numerical variables (means and standard deviations for each group). The *p*-values from a *t*-test are also reported. As it is evident, all the variables are significant at 1% level of significance, except the ratio inventories/cost of sales.For the significance of the binary attributes regarding the business activity of the firms was tested with a χ^2 test and all three indicators were found significant at the 1% level.

Table 3. Descripted statistics for the numerical attributes (training sample)

	Non-defaulted		Defaulted		
Variables	Mean	SD	Mean	SD	*p*-value
STL/CS	424.012	365.509	618.821	460.722	0.000
AR/S	210.207	200.712	343.483	337.496	0.000
I/CS	168.433	243.625	224.000	334.918	0.148
PBT/TA	0.043	0.132	–0.037	0.156	0.000
FE/S	0.026	0.034	0.052	0.054	0.000
QA/STL	1.123	0.854	0.847	0.613	0.000
TL/TA	0.726	0.265	0.854	0.264	0.000
LOGE	1.009	0.559	0.492	0.495	0.000

The discriminating power of the selected predictor attributes is also tested with a non-parametric test, namely the area under the receiver operating characteristic curve (AUROC). The results presented in Table 4 confirm the aforementioned remarks on the discriminating power of the variables.

Table 4. Area under the receiver operating characteristic curve (AUROC)

Variables	AUROC	*p*-value
STL/CS	0.374	0.000
AR/S	0.398	0.000
I/CS	0.503	0.923
PBT/TA	0.718	0.000
FE / SAL	0.373	0.000
FE/S	0.634	0.000
QA/STL	0.351	0.000
TL/TA	0.748	0.000
EXP	0.586	0.002
IMP	0.660	0.000
REPR	0.586	0.002

4. RESULTS

4.1. Setting of the Empirical Analysis

The analysis is performed under three different settings. In particular, first a full sample analysis is undertaken, where the full training data are used to build classification models (general models) with the selected methods. The models are then compared on the basis of the training data.

Through the second setting, we analyze the impact of using training samples of varying size on the performance and stability of the models. This analysis is performed by constructing different sets of 100 bootstrap samples with different proportions of defaulted/non-defaulted observations (adjusted model). Given the major imbalance in the size of the two classes in the training sample, we construct different bootstrap samples for the two groups. The observations in the samples corresponding to the default group are selected at random with replacement from the default cases in the training sample.

These bootstrap samples have size equal to the number of default cases in the training sample (i.e., 105). On the other hand, we vary the size of the bootstrap samples corresponding to non-default group, beginning with n_D observations up to n_{ND} (i.e., with $2n_D$, $5n_D$ and $10n_D$ serving as the intermediate cases), where n_D and n_{ND} denote the number of default and non-default observations, respectively, in the full training sample (n_D=105, n_D=5594). Thus, we begin with small-size bootstrap samples in which the two classes are fully balanced (i.e., 105 observations from each group), and gradually increase the size of the bootstrap samples, holding the number of default observations fixed, but increasing the number of the non-default.

Finally, in the third setting considered in the analysis, we examine the aggregation of results developed in the previous setting through the bagging approach (Breiman, 1996) in order to test the usefulness and predictive power of ensemble models.

The assessment of the predictive performance of all models is done using two metrics. The first one involves the accuracy rates, which show the ability of the models to correctly classify the firms. In particular, we consider the accuracy rates for each group of firms, the mean accuracy (the mean of the accuracies for each group) and the overall accuracy (i.e., the ratio of the correct predictions to the total observations in the testing sample). Additionally, the area under the receiver operating characteristic curve is also used, which enables the analysis of the predictive performance of classification rules under different hypotheses with respect to the misclassification costs and the a priory class membership probabilities (Fawcett, 2006).

4.2. Classification Results

Tables 5–7 summarize the results of the analysis. In particular, Table 5 presents the accuracy rates for each group of firms. Panels A and B present the accuracy rates for the two groups of firms, obtained with the models developed through each method for different sizes of the bootstrap samples. The results obtained the models fitted on the full training sample are also reported in the last column. The accuracies range between 75–80% in most of the cases for the non-default group, whereas the accuracies for the default class are lower, with the exception of the models developed with the UTADIS method.Overall, the UTADIS method achieves the best balance between the two groups, followed by the non-linear SVM model with the RBF kernel. Furthermore, it is worth observing that increasing the size of the bootstrap samples leads to higher estimates for the accuracy rates for the non-default group, whereas the predictions for the default class are almost unaffected. Compared to the results of the models fitted on the full sample, the bootstrap results tend to overestimate the accuracy rates for the non-default group (as the size of the bootstrap samples increases), whereas the accuracy rates for the default group are consistently underestimated.

Panels C and D in Table 5 present the accuracy rates for the two groups of firms for the ensemble models. Compared to the full models, it is evident that the ensembles perform better for the non-default group (especially as the size of the bootstrap samples increases), but worse for the default class. However, the differences are generally small. In fact, even with the smaller bootstrap samples satisfactory results can be obtained.

Table 6 summarizes the results for the mean and overall accuracy of all models. The LR and linear SVM model perform best on the basis of their overall accuracy, but this is due to their much better performance for the non-default group (which is much larger compared to the default class). On the other hand, the more balanced performance of the UTADIS models leads to improved average accuracy which is much higher compared to the other methods. The performance of the individual models built with bootstrap samples of small size is lower compared to the models developed using the full sample, but the results improve as the size of the bootstrap samples increases. Nevertheless, using the bootstrap analysis results to build ensemble models leads to results which are comparable and in some cases slightly bettercompared to the models developed through the full sample, even for bootstrap sample of small size. It should be noted that even through in terms of their overall accuracy the ensemble models seem to outperform the models developed through the full training sample, the average accuracy suggests that the improvement is marginal at best (it is slightly higher when small bootstrap samples are used).

Table 5. Accuracies rates (in %) for each group of observations

	Bootstrap samples size					Full sample
	205	315	630	1,155	5,699	
Panel A: Non-default group - individual models						
LR	74.92	76.85	78.36	78.57	79.51	78.50
SVM	74.47	75.89	76.69	76.90	78.07	77.39
SVM RBF	74.38	75.96	76.94	77.30	78.49	76.77
UTADIS	72.60	74.48	75.22	75.50	76.28	75.61
Panel B: Default group - individual models						
LR	67.61	66.73	67.36	67.50	66.09	67.80
SVM	69.45	69.06	69.59	69.96	68.38	70.63
SVM RBF	68.95	68.38	69.11	69.29	67.74	72.03
UTADIS	74.21	73.85	74.29	74.24	74.09	78.32
Panel C: Non-default group - ensemble models						
LR	77.39	77.94	79.26	79.13	80.00	
SVM	76.53	77.39	77.60	77.80	78.83	
SVM RBF	76.49	77.45	77.76	77.86	79.01	
UTADIS	74.68	76.32	76.32	76.51	77.35	
Panel D: Default group- ensemble models						
LR	69.93	67.83	67.13	67.83	65.73	
SVM	73.43	70.63	72.03	70.63	69.23	
SVM RBF	72.73	70.63	71.33	70.63	69.23	
UTADIS	78.32	79.02	77.62	77.62	76.22	

Table 6. Mean and overall accuracies

	Bootstrap samples size					Full sample
	205	315	630	1155	5699	
Panel A: Mean accuracy of individual models						
LR	71.27	71.79	72.86	73.04	72.80	73.15
SVM	71.96	72.47	73.14	73.43	73.23	74.01
SVM RBF	71.67	72.17	73.02	73.30	73.11	74.40
UTADIS	73.41	74.16	74.75	74.87	75.18	76.96
Panel B: Overall accuracy of individual models						
LR	74.72	76.56	78.05	78.26	79.13	78.20
SVM	74.32	75.69	76.49	76.70	77.79	77.20
SVM RBF	74.23	75.74	76.71	77.08	78.18	76.64
UTADIS	72.65	74.46	75.19	75.46	76.21	75.68

Table 6. (Continued)

	Bootstrap samples size					Full sample
Panel C: Mean accuracy of ensemble models						
LR	73.66	72.89	73.20	73.48	72.87	
SVM	74.98	74.01	74.81	74.21	74.03	
SVM RBF	74.61	74.04	74.54	74.25	74.12	
UTADIS	76.50	77.67	76.97	77.07	76.79	
Panel D: Overall accuracy of ensemble models						
LR	77.18	77.66	78.91	78.81	79.59	
SVM	76.44	77.20	77.44	77.60	78.55	
SVM RBF	76.38	77.26	77.58	77.66	78.73	
UTADIS	74.79	76.40	76.36	76.54	77.32	

The conclusions drawn from the above comparisons are further confirmed by the results shown in Table 7 on the AUROC. Again UTADIS provides the best results among the methods considered in the comparison and the ensemble models are only marginally better compared to the ones fitted on the full sample. Furthermore, the ensembles' results are not significantly affected by the size of the bootstrap samples and the associated imbalance in the size of the two classes.

Table 7. Area under the receiver operating characteristic curve

	Bootstrap samples size					Full sample
	205	315	630	1155	5699	
Panel A: Individual models						
LR	0.776	0.789	0.796	0.799	0.798	0.802
SVM	0.788	0.791	0.798	0.802	0.801	0.809
SVM RBF	0.783	0.788	0.797	0.802	0.801	0.812
UTADIS	0.800	0.807	0.813	0.814	0.816	0.824
Panel B: Ensemble models						
LR	0.807	0.799	0.809	0.810	0.806	
SVM	0.813	0.806	0.810	0.813	0.810	
SVM RBF	0.812	0.805	0.811	0.813	0.810	
UTADIS	0.824	0.826	0.828	0.829	0.828	

5. Conclusion and future Perspectives

This study presented an empirical comparison of three popular techniques for constructing credit risk assessment models using a large sample of more than 10,000 observations involving Greek firms from the commercial sector, over the period 2006–2009. The comparative analysis focused on different settings for the handling the training data in order to analyze the effect of the considerable class size imbalance on the predictive performance of the models. A bootstrap sampling approach was employed for this purpose and the construction of ensemble models through the bagging approach was tested.

Overall, the results showed that the UTADIS a multicriteria method performed better than logistic regression and support vector machines. Using training data of different size, it was observed that even with small samples good results can be obtained, which are improved as the sample size increases. On the other hand, an ensemble approach leads to results similar or slightly better compared to the full models, even with bootstrap samples of small size. This finding may have significant implications as far the computational aspects of the model development process is concerned, mainly for methods that do not scale well with the size of the training data.

Future research can be extendedtowards several directions. First, there should be a consideration of other methods (neural networks, multicriteria methods, rule-based models). Furthermore, other modeling approaches such as hazard models and survival analysis or option-pricing models could also be applied. Finally, it would be worthwhile considering additional predictors including among others economic conditions, stock market data and qualitative aspects (e.g., management's performance).

REFERENCES

Ahn, B.A., Cho, S.S., and Kim, C.Y. (2000),"The integrated methodology of rough set theory and artificial neural network for business failure prediction", *Expert Systems with Applications* 18,65–74.

Altman, E.I. (1968), "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *Journal of Finance* 13, 589–609.

Altman, E.I., Haldeman, R., and Narayanan, P. (1977), "ZETA^{TM} analysis: A new model to identify bankruptcy risk of corporations", *Journal of Banking and Finance* 1(1), 29–54.

Altman, E.I., Marco, G., and Varetto, F. (1994), "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (The Italian experience)", *Journal of Banking and Finance* 18(3), 505–529.

Altman, E.I., and Narayanan, P. (1997), "An international survey of business failure classification models", *Financial Markets, Institutions and Instruments* 6(2), 1–57.

Altman, E.I.and Saunders A. (1998), "Credit risk measurement: Developments over the last 20 years", *Journal of Banking and Finance* 21(11–12), 1721–1742.

Angelini, E., Di Tollo,G., andRoli, A. (2008),"A neural network approach for credit risk evaluation",*The Quarterly Review of Economics and Finance* 48,733–755.

Atiya, A. (2001), "Bankruptcy prediction for credit risk using neural networks: A survey and new results", *IEEE Transactions on Neural Networks* 12(4), 929–935.

Baesens, B., Setiono, R., Mues, C., and Vanthienen, J. (2003),"Using neural network rule extraction and decision tables for credit-risk evaluation", *Management Science* 79(3), 312–329.

Barnes, P. (1990), "The prediction of takeover targets in the UK by means of multiple discriminant analysis", *Journal of Business Finance and Accounting* 17(1), 73–84.

Beaver, R. (1966), "Financial ratios as predictors of failure", *Journal of Accounting Research* 4, 71–111.

Becchetti, L. and Sierra, J. (2003), "Bankruptcy risk and productive efficiency in manufacturing firms", *Journal of Banking and Finance* 27(11), 2099–2120.

BIS (2004), *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*, Basel Committee of Banking Supervision, Bank for International Settlements.

Bonfim, D. (2009), "Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics", *Journal of Banking and Finance* 33(2), 281–299.

Breiman, L.(1996), "Bagging predictors", *Machine Learning* 24,123–140.

Bryant, S.M. (1997), "A case-based reasoning approach to bankruptcy prediction modeling", *Intelligent Systems in Accounting, Finance and Management* 6,195–214.

Bugera, V., Konno, H., and Uryasev, S. (2002), "Credit cards scoring with quadraticutility functions", *Journal of Multi-Criteria Decision Analysis* 11(4–5), 197–211.

Cielen, A., Peeters, L., and Vanhoof, K. (2004), "Bankruptcy prediction using a data envelopment analysis",*European Journal of Operational Research* 154(2), 526–532.

Dimitras, A.I., Slowinski, R., Susmaga, R., and Zopounidis, C., (1999), "Business failure prediction using rough sets", *European Journal of Operational Research* 114(2), 263–290.

Doumpos, M., Kosmidou, K., Baourakis, G., and Zopounidis, C. (2002), "Credit risk assessment using a multicriteria hierarchical discrimination approach: A comparative analysis", *European Journal of Operational Research* 138(2), 392–412.

Doumpos, M. and Zopoounidis. C. (2002), *Multicriteria Decision Aid Classification Methods*, Kluwer Academic Publishers, Dordecht.

Doumpos, M. and Zopounidis, C. (2011), "A multicriteria outranking modeling approach for credit rating", *Decision Sciences* 42(3), 721–742.

Emel, A.B., Oral, M., Reisman, A., and Yolalan, R. (2003), "A credit scoring approach for the commercial banking sector", *Socio-Economic Planning Sciences* 37, 103–123.

Espahbodi, H. and Espahbodi, P. (2003), "Binary choice models for corporate takeover", *Journal of Banking andFinance* 27, 549–574.

Fawcett, T. (2006), "An introduction to ROC analysis", *Pattern Recognition Letters* 27(8), 861–874.

Gaganis. C., Pasiouras, F., Spathis, C., and Zopounidis, C. (2007), "A comparison of nearest neighbours, discriminant and logit models for auditing decisions", *Intelligent Systems in Accounting, Finance and Management* 15, 23–40.

Greco, S., Matarazzo, B., and Slowinski, R. (1998), "A new rough set approach to evaluation of bankruptcy risk", in: C. Zopounidis (ed.), *Operational Tools in the Management of Financial Risks*, Kluwer Academic Publishers, Dordrecht, 121–136.

Hamer, M.M. (1983), "Failure prediction: sensitivity of classification accuracy to alternative statistical methods and variable sets", *Journal of Accounting and Public Policy* 2, 289–307.

Hamerle, A.H., Liebig, T.L., and Scheule, H. (2006), "Forecasting credit event frequency - empirical evidence for West German firms", *Journal of Risk* 9(1), 75–98.

Hu, Y.-C. (2009), "Bankruptcy prediction using ELECTRE-based single-layer perceptron", *Neurocomputing* 72(13–15), 3150–3157.

Huang, Z.,Chen, H.C., Hsu, C.J., Chen W.H., and Wu, S.S. (2004),"Credit rating analysis with support vector machines and neural networks: A market comparative study", *Decision Support Systems* 37(4),543–558.

Karels, G.V. and Prakash, A.J. (1987), "Multivariate normality and forecasting of business bankruptcy", *Journal of Business Finance and Accounting* 14(4), 573–593.

Ko,L.J. Blocher E.J., and Lin,P.P. (2007), "Prediction of corporate financial distress: An application of the composite rule induction system", *The International Journal of Digital Accounting Research*1(1), 69–85.

Kocagil, A.E.,Escott, Ph.,Glormann, F., Malzkorn, W., and Scott, A.(2002), "Moody's RiskCalc for private companies: UK", Moody's Investors Service, Global Credit Research.

Kolari, J., Glennon, D., Shin, H., and Caputo, M. (2002), "Predicting large US commercial bank failures", *Journal of Economics and Business* 54(4), 361–387.

Kou, G., Peng, Y., Shi, Y., Wise, M., and Xu, W. (2005), "Discovering creditcardholders' behavior by multiple criteria linear programming", *Annals of Operations Research* 135(1), 261–274.

Lacher, R.C., Coats, P.K., Sharma, S.C., and Fantc, L.F. (1995), "A neural network for classifying the financial health of a firm", *European Journal of Operational Research* 85(1),53–65.

Lee, K.C., Han, I., and Kwon, Y. (1996), "Hybrid neural network models for bankruptcy predictions", *Decision Support Systems* 18,63–72.

Leshno, M. and Spector, Y. (1996), "Neural network prediction analysis: The bankruptcy case", *Neurocomputing* 10, 125–147.

Lopez, J.A. and Saidenberg, M.R. (2000), "Evaluating credit risk models", *Journal of Banking andFinance* 24(1), 151–165.

McKee, T.E. (2003), "Rough sets bankruptcy prediction models versus auditor signaling rates", *Journal of Forecasting* 22,569–589.

McKinsey (1993), Special Report on the "New world of financial services", The McKinsey Quarterly, Number 2.

Min, J.H. and Lee, Y.-C. (2005), "Bankruptcy prediction using support vector machine (SVM) with optimal choice of kernel function parameters", *Expert Systems with Applications* 28,603–614.

Ohlson, J. (1980), "Financial ratios and the probabilistic prediction of bankruptcy",*Journal of Accounting Research* 18, 109–131.

Palepu, K.G. (1986), "Predicting takeover targets: A methodological and empirical analysis", *Journal of Accounting and Economics*8, 3–35.

Park, C.-S. and Han, I. (2002), "A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction", *Expert Systems with Applications* 23(3),255–264.

Piramuthu, S., Ragavan, H., andShaw, M.J. (1998), "Using feature construction to improve the performance of the neural networks", *Management Science* 44(3), 416–430.

Piramuthu, S. (1999), "Financial credit-risk evaluation with neural and neurofuzzy systems", *European Journal of Operational Research* 112(2), 310–321.

Platt, H.D. and Platt, M.B. (1991),"A note on the use of industry-relative ratios in bankruptcy prediction", *Journal of Banking andFinance*15, 1183–1194.

Schölkopf, B. and Smola, A. (2002), *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*,MIT Press,Cambridge.

Shin, K.-S. and Lee, Y.-J. (2002), "A genetic algorithm application in bankruptcy prediction modeling", *Expert Systems with Applications* 23(3), 321–328.

Smith, L.D. and Lawrence, E. (1995), "Forecasting losses on a liquidating long-term loan portfolio", *Journal of Banking and Finance* 19, 959–985.

Stecking, R. and Schebesch, K.B. (2003), "Support vector machines for credit scoring: Comparing to and combining with some traditional classification methods", in Schader, M., Gaul, W., and Vichy, M. (eds.), *Between Data Science and Applied Data Analysis*, Springer-Verlag, Berlin, 604–612.

Swicegood, P. and Clark, J.A. (2001), "Off-site monitoring for predicting bank under performance: A comparison of neural networks, discriminant analysis and professional human judgment", *International Journal of Intelligent Systems in Accounting, Finance and Management* 10,169–186.

Troutt, M.D., Rai, A., and Zhang, A. (1996), "The potential use of DEA for credit applicant acceptance systems", *Computers and Operations Research* 23(4), 405–408.

Tung, W.L., Quek, C., and Cheng, P. (2004), "GenSo-EWS: A novel neural-fuzzy based early warning system for predicting bank failures", *Neural Networks* 17, 567–587.

Vapnik, V. (1998), *Statistical Learning Theory*, Wiley, New York.

West, D., Dellana, S., and Qian, J. (2005), "Neural network ensemble strategies for financial decision applications", *Computers and Operations Research* 32, 2543–2559.

Yurdakul, M. and Tansel, Y. (2004), "AHP approach in the credit evaluation of the manufacturing firms in Turkey", *International Journal of Production Economics* 88, 269–289.

Zopounidis, C. and Doumpos. M. (1999), "Business failure prediction using the UTADIS multicriteria analysis method", *Journal of Operational Research Society*50(11), 1138–1148.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.