REVISTA ESPAÑOLA DE FINANCIACIÓN Y CONTABILIDAD
Vol. XXXVII, n.º 137 · enero-marzo 2008 · pp. 13-32

Linear discriminant analysis versus adaboost for failure forecasting*

Discriminante lineal versus adaboost para la predicción del fracaso empresarial

Esteban Alfaro Cortés. Universidad de Castilla-La Mancha

Matías Gámez Martínez **. Universidad de Castilla-La Mancha

Noelia García Rubio. Universidad de Castilla-La Mancha

ABSTRACT Since the sixties, many classification techniques both from statistics and other scientific disciplines have been used to predict corporate failure. Multivariate linear discriminant analysis, however, continues to be one of the main reference methods principally because it is easy to apply and interpret. AdaBoost is a machine learning technique which works by combining a large number of simple classifiers to achieve a high level of accuracy in classification problems. Although the efficiency of this method has been proved in various application fields, it is unknown in the economic-business area. The aim of this study is to present AdaBoost as a classification technique which can be successfully used in failure forecasting. Correspondingly, it has been applied to a sample of 1.180 Spanish firms and has proved to be more accurate than discriminant analysis.

KEYWORDS Corporate failure forecast; Discriminant analysis; Adaboost.

RESUMEN Desde la década de los sesenta, han sido muchas las técnicas de clasificación, provenientes tanto de la Estadística como de otras disciplinas científicas, que se han utilizado para la predicción del fracaso empresarial. Sin embargo, el análisis discriminante lineal multivariante sigue siendo una de las técnicas de referencia, fundamentalmente por su fácil aplicación e interpretación.

Adaboost es una técnica procedente del aprendizaje automático, que actúa combinando gran cantidad de clasificadores sencillos para conseguir un elevado grado de precisión en problemas de clasificación. La eficiencia de este método ya ha sido probada en diversos campos de aplicación, sin embargo, es desconocido en el ámbito económico-empresarial.

El objetivo de este artículo es presentar Adaboost como técnica de clasificación que se puede utilizar con éxito en la predicción del fracaso empresarial. Para ello se ha aplicado a una muestra de 1.180 empresas españolas superando en precisión al análisis discriminante.

PALABRAS CLAVE Predicción del fracaso empresarial; Análisis discriminante; Adaboost.

* We would like to thank Domingo García Pérez de Lema for his help in revising this article. We would also like to thank Leandro Cañibano Calvo and Gregorio Labatut Serer for their comments within the framework of the V Workshop de Investigación Empírica en Contabilidad Financiera, and two anonymous reviewers and the editor for their observations during the revision process.

** Correspondence address: Matías Gámez Martínez, Departamento de Economía Política y Hacienda Pública, área de Estadística Económica y Empresarial, Plaza dela Universidad 1, 02071 Albacete. Tel. 967599200, ext. 2331, E-mail: Matias.Gamez@uclm.es.

Recibido 11-07-06 · Aceptado 04-06-07 · · ISSN: 0210-2412

1. INTRODUCTION

Corporate failure prediction consists in separating the firms with a high probability of future failure from those which are considered to be healthy. In order to do so, it is necessary to know the financial state of the firms. This is a matter which has been studied for almost four decades and is still one which concerns the scientific community, and while there have been many methodological improvements during this time, these have not always implied an improvement in results. Nowadays, the subject is attracting the attention of a great many researchers due to the availability of financial information which was very hard to find not so many years ago. Another factor that has contributed to the increasing interest is the use of alternative analysis tools from machine learning. Moreover, qualitative attributes are being tested which when combined with traditional numerical ones improve the accuracy of the forecasting models and knowledge of the factors which determine whether a firm will survive or not.

According to Gallego and Gómez (2002), two main research lines can be distinguished among the numerous studies into failure forecasting: on the one side, that which focuses on finding classification methods with better prediction accuracy, and on the other, that which focuses on the most relevant ratios for minimizing prediction error. Unfortunately, experts do not agree on the best ratios to select to solve the problem. This article is encompassed by the first line since it proposes a novel method which is well known in the field of artificial intelligence due to its high levels of accuracy with an improvement on the results provided by other more traditional methods.

Since Beaver's pioneering work (1966) that solved the failure prediction task using financial ratios in a univariate way, a great number of papers have attempted to improve on the results. Other pioneering work was carried out by Altman (1968) with the application of multivariate discriminant analysis. Even today, this model still constitutes the reference point for comparing the results of new contributions.

After this pioneering research, other models such as conditional probability, logit (Ohlson, 1980) and probit (Zmijewski, 1984) have been applied to avoid the very restrictive hypothesis regarding normality and equality of variance-covariance matrices. More recently, classification trees (Frydman *et al.,* 1985) and (Gabás, 1990) or artificial neural networks (Wilson and Sharda, 1994) and (Serrano Cinca, 1997) have been proposed as alternative methods for failure prediction.

As mentioned previously, the comparison of different classification methods is a very important research line that constitutes the goal of many papers. In Spain, we can mention the work by Serrano Cinca (1997) that compares discriminant analysis and artificial neural networks by means of the leaving-one-out cross-validation method using a sample of 66 Spanish banks (29 failed and 37 healthy). By way of conclusion, the neural model performs better with an accuracy of 93.94%, 7.58 points above the 86.36% reached by discriminant analysis.

Ferrando and Blanco (1998) compare the discriminant and logit model abilities for failure prediction in the Comunidad Valenciana. As we have in our work, they use legal failure as surrogate failure working with companies in receivership or bankrupt in the 1992-1994 period. The sample consisted of 88 failed firms and another 88 healthy ones. The total sample is divided into two sets: the training data set with 120 cases and the validation set with 56 cases, maintaining the parity between classes. The discriminant function obtains an accuracy of 87.5% over the validation data set. On the other hand, the logit model obtains an accuracy of 85.71% over the validation sample taking into account the information one year before the failure process. Both models, however, obtain the same results if the information is taken for the two years prior to failure. The authors conclude that the logit model is slightly better than discriminant analysis on the basis of the Type I and Type II errors and the percentage of correctly classified cases in the estimation samples. Nevertheless, Ferrando and Blanco do not study the statistic significance of the differences, which can be carried out using any resampling technique.

At a later stage, Rodríguez López (2001a) provided additional empirical evidence of the prediction of failure for small and medium non-financial firms in Galicia. The concept of failure used in this work is wider than the legal concept of bankrupt firms or those in receivership and also includes those firms involved in judicial processes involving debt collection or unpaid bills of exchange figuring in registries elaborated by specialized companies, all with a high number and amount. The sample consists of 120 firms, half of which had failed and the other half were healthy. The validation sample comprised 29 failed firms and 284 healthy ones. Discriminant and logit analysis were the techniques used. One model is estimated for each of the four years prior to the moment of failure and an overall one is estimated from the information as a whole. The total percentages of cases correctly classified by the discriminant model over the validation sample are (from closer to farther from the time of failure) 97.4, 88.5, 73.2, 71.6% and 7.9% for the overall model. The results achieved by the logit model are 96.5, 84.6, 80.8 and 69.2%, respectively. The logit model performs better in the cases of the overall model and the three years before the failure model.

On an international level, we can mention the work by Charalambous *et al.* (2000) that applies several versions of artificial neural networks to a set of 139 pairs of American failed and healthy firms for the period 1983-1994. More specifically, these authors compare the generalization ability of six methods, radial basis function, feedforward nets using the conjugate gradient algorithm, the LVQ (Learning Vector Quantization) method, back-propagation algorithm, self-organizing feature maps (SOFM) proposed by Kohonen and logistic regression, with the feed-forward networks and the SOFM obtaining the best results with accuracies for the test data sets of 82.6 and 80.2%, respectively. Other interesting recent publications include Pérez (2006) and Ravi Kumar and Ravi (2007), which collect the studies relating to failure prediction using artificial neural networks in the first case and several statistical techniques in addition to methods from artificial intelligence in the second.

In line with Labatut *et al.* (2006) and Abad et al. (2004), we believe that the new Basel II regulatory framework has increased the importance of finding extremely precise models to predict the failure process. This agreement represents a significant change in relation to the Basel I Capital Accord (1988), and its subsequent modification in 1996, since it allows financial entities to calculate their minimum capital requirements for credit risk *through their own accredited classification systems and the measure of risk factors associated to each category*. For this, financial entities will need to demonstrate to supervisors that they have internal procedures to evaluate the adequacy of their capital in relation to their risk profile. This new approach involves important challenges for both entities and supervisors. The financial entities that want to use their own models will need to be prepared to comply with the established requirements. More specifically, they will have to evaluate whether their measuring systems, databases and procedures are suitable for complying with such requirements. On the other hand, supervisors will have to be ready to validate the models used to calculate the minimum capital requirements and decide whether the entity's capital is sufficient for its risk profile.

Due to the increasing importance given to the accuracy of the models, in this paper we propose a novel approach as an alternative for failure prediction that enables us to obtain higher levels of precision and this model is known as AdaBoost. Various publications exist which offer a thorough comparison of this method with other classification systems (artificial neural networks, linear discriminant analysis, logit, classification tress, etc.) using certain well-known databases such as UCI (University of California, Irvine). For instance, the work by Bauer and Kohavi (1999) and Breiman (1998) should be mentioned. The results have shown that AdaBoost performs better than single classifiers and it is one of the best ensemble classifiers. We have chosen the discriminant analysis as a reference point for comparison since it is one of the most popular techniques in the field of economics. In addition, Rodríguez (2001b) and Ferrando and Blanco (1998) defend the robustness of discriminant analysis over deviations from the hypotheses on which they are based. Moreover, as mentioned previously in Rodríguez's work (2001a), discriminant analysis behaves better than the logit model for one, two and four years before the moment of failure. Discriminant analysis is therefore the winner in three out of five applications and precisely in the time periods in which we work.

In addition to the AdaBoost method, we propose a measure to discover the influence or importance of a variable to facilitate model interpretation. In this way, it is possible to discover the most important variables when it comes to discriminating between failed and healthy firms.

The work is structured as follows. In the next section we present the AdaBoost algorithm and briefly explain how it should be applied [(1)]. Section 3 introduces the sample characteristics for the empirical application. In Section 4, we show the results obtained with both discriminant and AdaBoost methods. The results obtained by the two methods are compared both for one year and two years before failure. Finally, the main conclusions are drawn in Section 5.

2. ADABOOST ALGORITHM

A classifier system builds a model which is able to predict the class of a new observation given a data set. The accuracy of the classifier will depend on the quality of the method used and the difficulty of the specific application. If the obtained classifier achieves a better accuracy than the default rule, then the classification method has found some structure in the data enabling it to do so. AdaBoost is a method that makes maximum use of a classifier by improving its accuracy. The classifier method is therefore used as a subroutine to build an extremely accurate classifier based on the training set.

AdaBoost applies the classification system repeatedly to the training data, but at each application, the learning attention is focused on different examples of this set using adaptive

(1) We do not consider it necessary to explain the discriminant analysis technique because it is a classical and well-known method (Fisher, 1936), (Hand, 1981), (Uriel, 1995).

weights [$\omega_b(i)$]. Once the training process has finished, the single classifiers obtained are combined into a final, highly accurate classifier based on the training set. The final classifier therefore usually achieves a high degree of accuracy in the test set as various authors have shown both theoretically and empirically, among which we must cite Bauer and Kohavi (1999), Breiman (1998), Drucker and Cortes (1995), Dietterich (2000), Friedman, Hastie and Tibshirani (2000), Freund and Schapire (1996, 1997) and Schapire (2002).

Even though there are several versions of boosting algorithms, the most widely used is the one by Freund and Schapire which is known as AdaBoost. For simplification purposes, it can be assumed that there are only two classes without loss of generality. A training set is given by T = {(x_1^b, y_1^b), ..., (x_n^b, y_n^b)}where Y takes values of {-1, 1}. The weight $\omega_b(i)$ is assigned to each observation X_i and is initially set to $1/n$. This value will be updated after each step. A basic classifier denoted $C_b(X_i)$ is built on this new training set, T^b, and is applied to each training sample. The error of this classifier is represented by ε_b and is calculated as

$$\varepsilon_b = \sum_{i=1}^{n} \omega_b(i)\xi_b(i) \qquad donde \quad \xi_b(i) = \begin{cases} 0 & C_b(x_i) = y_i \\ 1 & C_b(x_i) \neq y_i \end{cases} \tag{1}$$

The new weight for the (b+1)-th iteration will be

$$\omega_{b+1}(i) = \omega_b(i)\, exp[\alpha_b \xi_b(i)] \tag{2}$$

where α_b is a constant calculated from the error of the classifier in the b-th iteration. More specifically, according to the authors mentioned above $\alpha_b = ln[(1 - \varepsilon_b)/ \varepsilon_b)$. The calculated weights are then normalized so that they add up to one. Accordingly, $\varepsilon_b = 0.5 - \gamma_b$, where γ_b shows the advantage of the basic classifier of the *b*-th step over the default rule in the worst case, where both classes have the same *a priori* probability (0.5).

The following example shows how the weights are updated as a function of the error in the second iteration, *b* = 2. For this purpose, we have selected two values close to the limits of the range of ε_b: i.e. 0 and 0.5. If $\varepsilon_b = 0.499$, then $\alpha_b = 0.004$ and the new weight for the second iteration will be $\omega_2(i) = 1/n$ exp (0.004 $\xi_b(i)$). Then, if the i-th observation is wrongly classified, its weight will be $\omega_2(i) = 1/n \cdot 1.004$, whereas if it is correctly classified, its weight remains the same although it will be decreased due to the normalization process.

If we now consider that $\varepsilon_b = 0.001$, then $\alpha_b = 6.907$ and the weight of a wrongly classified observation will be ω_2(i)=$1/n \cdot 999$, while the weights of correctly classified observation will be reduced by normalization. Table 1 shows the updated and normalized weights after the second iteration for a set of 1000 observations.

TABLE 1
EXAMPLE OF HOW WEIGHTS ARE UPDATED IN ADABOOST

n	*initial weight*	*error*	*alpha*	*wrong class.*	*weight 2*	*norm. weight 2*
1000	0.001	0.499	0.004	1	0.001004008	0.001002004
1000	0.001	0.499	0.004	0	0.001	0.000998004
1000	0.001	0.001	6.907	1	0.999	0.5
1000	0.001	0.001	6.907	0	0.001	0.000500501

It can be seen how the weights of the wrongly classified observations are increased and the weights of the correctly classified ones are decreased, forcing the single classifier built in the following iteration to focus on the hardest examples. Furthermore, the smaller the error made by the single classifier, the greater the differences when the weights are updated, since greater importance is given to the few mistakes made when the classifier achieves a high level of accuracy. The alpha constant can therefore be interpreted as a learning rate which is calculated as a function of the error made in this iteration. This constant is also used in the final decision rule with more importance being given to the individual classifiers that made the smallest error.

This process is repeated at each step for $b = 1, 2, 3, \dots, B$. Finally, the ensemble classifier is built as a linear combination of the single classifiers weighted according to the corresponding constant α_b.

$$C(x) = sign\left(\sum_{b=1}^{B} \alpha_b C_b(x)\right) \quad (3)$$

AdaBoost can be applied in two ways, either by using resampling or reweighting. For the first version, a data set Sb is obtained for the b-th iteration from a bootstrap sample extracted with replacement and using the weight of each observation as the extraction probability. In the reweighting version, the classifier C_b takes the weight into account directly. There is not enough evidence about the superiority of one method over another as is apparent from the work by Breiman (1998), Freund anf Schapire (1997) and Freund and Schapire (1998).

The AdaBoost algorithm is shown below:

AdaBoost Algorithm (Freund and Schapire, 1996)
1. Start with $\omega_b(i) = 1/n$, $i = 1, 2, \dots, n$.   2. Repeat for $b = 1, 2, \dots, B$ *a)* Fit the classifier $C_b(x) \in \{-1,1\}$ using weights $w_b(i)$ on T^b.   *b)* Compute: and $\varepsilon_b = \sum_{i=1}^{n} \omega_b(i)\xi_b(i)$ and $\alpha_b = ln[(1 - \varepsilon_b)/\varepsilon_b]$ *c)* Update the weights $\omega_{b+1}(i) = \omega_b(i) \cdot exp[\alpha_b \xi_b(i)]$ and normalize them.   3. Output the final classifier $C(x) = sign\left(\sum_{b=1}^{B} \alpha_b C_b(x)\right)$

3. SAMPLE DESCRIPTION

The companies used in this study were selected from the SABI database of Bureau Van Dijk (BVD), one of Europe's leading publishers of electronic business information databases and provider of the Wharton Research Data Services. SABI covers all the companies whose accounts are placed on the Spanish Mercantile Registry. Since this database only collects information for commercial and industrial companies, financial companies are excluded from the sample and consequently from the analysis. We have approached the main task in this work as a dichotomised classification problem in an attempt to separate those firms which are likely to fail in the future from the healthy ones. Of the various available concepts of surrogate failure we have chosen the most popular in this kind of study: legal failure. According to this concept, failed companies are those that have gone into receivership or been declared bankrupt.

The reference period for the failed firms is the period 2000-2003. As in the majority of studies dealing with failure prediction, we work with companies that have gone into receivership or been declared bankrupt not in a single year but over a period of several years so that a higher number of failed firms can be included in the sample. From all the failed firms in the period, we only select those which have information about all the variables at the moment of failure and in the five previous years. The reason for this is that this work is part of wider research into attempting to predict failure up to five years before it happens (although in this work we focus exclusively on the year and two years before failure in order to clarify the presentation). The temporal distribution of receiverships and bankruptcies throughout the period 2000-2003 can be seen in Appendix 1. Information about the variables must be considered in relative terms to the year of failure, t, with t-1 and t-2 being the two previous years. On the other hand, the healthy firms were those that remained active at the end of 2003 and had complete information for 2003 and the five previous years. In such a case, there was an additional requirement: any firm with continual negative profits over the last three years would be rejected since even though they were still active in December 2003, they would soon enter a state of failure if they continued to make a loss. With these requirements, 590 failed firms were available and another healthy 590 firms were selected. We therefore had 1180 companies, 80% of which (944) were included in the training data set and the remaining 20% (236) formed the test set.

The variables or predictors have been chosen on the basis of two criteria: firstly, they should have been commonly used in failure prediction literature, and secondly, the information needed to calculate these ratios should be available. While there is no business failure theory to determine the ratios to be used, there are some commonly used ratios and these have been taken into account for this work. Thirteen financial ratios have therefore been selected (the ratios shown in Appendix 2 have previously been used in work by Abad et al. (2004), Frydman et al. (1985), Rodríguez (2001a) and Sanchís et al. (2003)). In addition, we decided to include two variables: the sector as a qualitative variable with ten categories using the National Classification of Economic Activities (NACE-93 digit-1 level), and the size using the natural logarithm of Total Assets as a proxy variable. The legal structure was also used as a categorical explanatory variable with three options: public corporations, limited corporations, and other corporations. The sample structure according to these three variables (activity sector, legal structure and size) can be seen in Appendix 3.

Sixteen predictor variables were therefore used for each of these Spanish companies. It must be stressed that although the sample size of each class is the same they are not paired by either the activity sector or the size.

There are various reasons behind the decision to work with the same size samples for both classes rather than maintaining the population proportion. Among these, Rodríguez (2001b) highlights the fact that the population proportion significantly favours active firms and so a random sampling process would result in a reduced number of failed firms in the sample. In this case, the information contained in the sample would be limited and therefore model estimation using the random procedure might lead to biased estimators.

Moreover, the true proportion of failed and healthy firms in the population is not easy to calculate in practice. Lastly, as Rodríguez noted (2001a), if the differences between the *a priori* probabilities were taken into account, the type I and II error costs would also need to be considered, i.e. the cost of classifying a healthy firm as failed and vice versa. Since the cost of classifying a failed firm as healthy is much higher than the cost of classifying a healthy firm as failed, the disproportions between *a priori* probabilities and error costs could be offset and produce an almost neutral effect.

As a result of all this, we decided to work with the same number of firms for both classes. However, as previously mentioned, the classes are not paired as they traditionally are in failure prediction work since in this way the size and activity effects are eliminated as potential discriminatory variables. In this paper, the firm size and the sector of activity are used as discriminatory variables in lnTA and NACE1.

4. EMPIRICAL RESULTS

We will now present the results obtained with Linear Discriminant Analysis (LDA) and AdaBoost, analysing first the case of one year before the moment of failure and then the results corresponding to the second year before.

4.1. First year before failure

We will first build the model with LDA. This method is well known so we will not dwell on its theoretical considerations. This method was proposed by Fisher in 1936 and a detailed explanation can be found in Uriel (1995). In order to apply the model, we have used the statistical package SPSS 13.0 for Windows and the objective variable has been coded with the values 0 for the failed firms and 1 for the healthy ones. We have also included the qualitative variables NACE1 (with 10 categories) and LEGAL STRUCTURE (with 3 categories) as dummy variables (1, 0) in order to create the same number of variables as the number of categories minus one for each attribute. We therefore need 9 variables for NACE1 and 2 for LEGAL STRUCTURE. In each dummy variable, we assign the value 1 to those firms that show the corresponding modality and 0 to the remaining firms. When working with a high number of explanatory variables, it is common practice to reduce their size using factorial analysis or alternatively to introduce the variables in the LDA model using the stepwise method. We have chosen the last option because it facilitates model interpretation by allowing us to work directly with the ratios rather than the factors. Additionally, the coherence of the obtained signs has been checked with those expected for each ratio. The step-by-step introduction method iteratively introduces the variable with the greatest discriminatory power measured using the Wilks' lambda statistic, and also enables any variable to be dropped if it becomes redundant or insignificant after any iteration. This is an iterative process that stops when the inclusion of a new variable does not significantly improve the model's discriminatory capacity. Table 2 summarizes the iterative process that has selected seven explanatory variables for the construction of the discriminant model.

TABLE 2
SUMMARY OF THE STEPWISE PROCESS

Step	Entered variable	Wilks' lambda	Signification level
1	S.TA	0,800	0,000
2	NACE_5	0,783	0,000
3	PUBLIC CORP.	0,773	0,000
4	CA.TA	0,766	0,000
5	EBIT.TA	0,738	0,000
6	lnTA	0,728	0,000
7	L.TD	0,723	0,000

Table 3 shows the unstandardized discriminant coefficients. These coefficients are the ones to be used to assign a class to each observation. The selected variables for the final model are profitability, weight of the current asset, indebtedness, efficiency, size of the firm, sector with code 5 in NACE *(wholesale and retail trade; repair of motor vehicles, motorcycles and personal/household goods)*, and finally, whether the firm is a public corporation or not. The firms which are most likely to fail are those with a high weight of current assets over total assets, those with a high ratio of indebtedness, and those which belong to NACE Sector 5. However, the firms which are most likely to remain active are those with higher profitability, those with a high efficiency ratio, larger firms, and those with the legal structure of a public corporation.

On the other hand, it is imperative to confirm that these variables have a significant joint capacity to discriminate between the two classes. For this, Wilks' lambda statistic is computed and in this case it yields a value of 0,723 as shown in Table 2. The probability of finding a more extreme value in the chi-squared distribution associated to this statistic is 0,000, which allows us to reject the null hypothesis of equality of mean vectors between the two classes.

TABLE 3
UNSTANDARDIZED COEFFICIENTS OF THE CANONICAL DISCRIMINANT FUNCTION

	Function 1
EBIT.TA	0,010
CA.TA	-0,097
L.TD	-0,016
S.TA	0,983
LnTA	0,164
NACE_5	-0,566
PUBLIC CORP.	0,286
(Constant)	-2,168

Using the discriminant coefficients and the values of the variables in each firm, we obtain the discriminant score for each case. On the basis of this score, each firm is assigned to one of the two groups (to the failed group if the score is lower than 0,5 and to the active group if the score is higher than 0,5). Repeating the process for the entire training and test data sets, we obtain the classification errors for each firm a year before failure. Table 4 shows the confusion matrix for both data sets and the corresponding errors. In this case, both errors are close to 20%: 20,763% in the training set and 20,339% in the test set. These error percentages are acceptable if we take into account that we have not paired according to sector or size and there is therefore a certain degree of heterogeneity in the sample which does however provide a wider level of applicability. The model therefore has an accuracy of 79.66% when it comes to predicting the future of new firms for the coming year. It is important to stress that for both sets the Type I error is higher than the Type II error.

TABLE 4

CONFUSION MATRICES AND ERRORS WITH LDA

<table>
<tr><td rowspan="5">PREDICTED CLASS</td><td colspan="5">ACTUAL CLASS</td></tr>
<tr><td rowspan="2">LDA</td><td>Training</td><td>20.763%</td><td>Test</td><td>20.339%</td></tr>
<tr><td>Failed</td><td>Healthy</td><td>Failed</td><td>Healthy</td></tr>
<tr><td>Failed</td><td>357</td><td>81</td><td>84</td><td>14</td></tr>
<tr><td>Healthy</td><td>115</td><td>391</td><td>34</td><td>104</td></tr>
</table>

Once we have analysed the results provided by the LDA model, we will compare them to the results yielded by the AdaBoost classifier. To do this, we have used the free statistical software R, and more specifically the available implementation in the R library adabag (Alfaro *et al.,* 2006). When the AdaBoost function is applied to decision trees, two parameters must be set: the number and size of the trees. The tree size, for its part, can be limited in several ways. Firstly, by fixing the maximum depth of the trees, i.e. the maximum number of nodes that an observation must pass through to reach a leaf node. Secondly, we can stop the growth of the tree by requiring a minimum number of observations in one node to make a new split. Finally, using the complexity parameter *(cp)* so that only the splits which reduce the impurity by a cp factor will be made.

After several initial experiments, we have decided to construct an AdaBoost classifier with one hundred trees grown to a maximum depth of 2. The error on the test set reaches 8.898%, which implies a reduction of 56.25% in comparison to the LDA error (20.339%). Moreover, it can be seen from the analysis of the confusion matrix how most errors come from the classification of healthy firms as failed while only four failed firms were classified as healthy and so Type I errors are lower with AdaBoost than Type II ones. Table 5 shows the confusion matrices and the error percentages for training and test sets.

TABLE 5

CONFUSION MATRICES AND ERRORS WITH ADABOOST

<table>
<tr><td rowspan="5">PREDICTED CLASS</td><td colspan="5">ACTUAL CLASS</td></tr>
<tr><td rowspan="2">ADABOOST</td><td>Training</td><td>7.627%</td><td>Test</td><td>8.898%</td></tr>
<tr><td>Failed</td><td>Healthy</td><td>Failed</td><td>Healthy</td></tr>
<tr><td>Failed</td><td>460</td><td>60</td><td>114</td><td>17</td></tr>
<tr><td>Healthy</td><td>12</td><td>412</td><td>4</td><td>101</td></tr>
</table>

In order to ensure that the comparison between the LDA model and the AdaBoost ensemble does not happen by chance we used five repetitions of 10-fold cross-validation. The entire set (1180 firms) is used for each 10-fold cross-validation experiment. We consequently obtained the error rates for the LDA and AdaBoost models on each of the 50 experiments. Having checked the normality of the error distributions of both classification methods using the tests shown in Table 6, we can apply the test for comparing the means of two normal distributions. To achieve a higher degree of certainty, we used the one tail test, establishing the null hypothesis that the average error of the AdaBoost classifier is equal to or higher than the LDA average error rather than the alternative hypothesis which establishes that it is lower. In this way, if we reject the null hypothesis, we will trust the alternative. The result obtained is enlightening because the t statistic is -20.1572 and its corresponding p-value is 0.000. We can therefore reject the null hypothesis and state that the differences are statistically significant with AdaBoost ahead.

TABLE 6

NORMALITY TESTS, AVERAGE AND STANDARD DEVIATION OF ERROR

Test	*LDA*	*ADABOOST*
Kolmogorov-Smirnov	0.1188	0.1141
KS. p value	0.4801	0.5327
Shapiro	0.9704	0.9375
Shapiro p value	0.2404	0.0107
Average error	0.2103	0.0956
Standard deviation	0.0241	0.0322

To facilitate model interpretation, we have programmed the AdaBoost function of the adabag package in such a way that it allows us to quantify the relative importance of the predictor variables. In this case, the most important ratios are EBIT.TA, L.TD, S.TA and lnTA with values at this measure of 17.33, 13.72, 10.47 and 9.02%, respectively. On the other hand, the least important ratios are CA.CL, S.CA and WC.S with values below 2%. Table 7 shows all the variables arranged from greatest to least relative importance.

The most important ratios in the analysis are economic profitability, indebtedness, efficiency, and the proxy variable for the size of the firm. Firstly, economic profitability shows the corporation's success in the application of assets, measuring these by means of the weight of generated earnings before interest and taxes. The most efficient firms in this aspect will undoubtedly have a greater likelihood of being classified as healthy. Secondly, the L.TD (Liabilities/Total Debt) ratio considers the weight of indebtedness in the financial structure, showing that firms with lower levels of indebtedness are more able to apply for new external financing sources and therefore have greater possibilities of survival. Thirdly, S.TA refers to the ratio of sales over total assets and shows, in a figurative sense, how many times the asset has been sold and, therefore, its turnover, i.e. how many times it has been sold and replaced. This ratio can also be interpreted as the sales in euros for each euro invested in assets, reflecting the capacity of assets to generate sales and the firm's relative efficiency in managing them. The higher this ratio (for the same commercial margin), the higher the profit for a lower investment, and therefore, the higher the profitability.

TABLE 7
RELATIVE IMPORTANCE OF VARIABLES

Variable	*Relative importance*	*Variable*	*Relative importance*
EBIT.TA	17.33	CA.TA	5.05
L.TD	13.72	EBIT.CAP	3.61
S.TA	10.47	LEGAL STR.	2.88
LnTA	9.02	S.CAP	2.17
C.CL	8.66	WC.TA	2.17
NACE_1	8.30	CA.CL	1.80
CF.TD	6.50	S.CA	1.80
C.TA	5.05	WC.S	1.44

4.2. SECOND YEAR BEFORE FAILURE

In this case, nine discriminatory variables are selected for the sequential process. A summary of the iterative process can be seen in Table 8.

TABLE 8
SUMMARY OF THE STEPWISE PROCESS (TWO YEARS BEFORE)

Step	*Entered variable*	*Wilks' lambda*	*Signification level*
1	LnTA	0.6472	0.000
2	LIMITED CORP.	0.6364	0.000
3	S.CAP	0.6289	0.000
4	S.CA	0.6215	0.000
5	L.TD	0.6172	0.000
6	NACE_4	0.6135	0.000
7	NACE_1	0.6087	0.000
8	EBIT.TA	0.6059	0.000
9	NACE_3	0.6031	0.000

Table 9 shows the unstandardized coefficients of the canonical discriminant function. The selected variables for the final discriminant model are profitability, size of the firm, indebtedness, rotation of capital in relation to level of sales, efficiency (sales over working capital), NACE Sectors 1, 3 and 4 and, finally, the legal structure. The firms most likely to fail are those with a high ratio of indebtedness and those which have a high level of rotation of the capital in relation to sales (both situations indicating an excessive use of indebtedness). However, the firms with the greatest chances of remaining healthy are highly profitable ones, larger firms, those belonging to Sectors 1, 3 and 4, and limited corporations. The joint discriminatory capacity of these variables is shown by Wilks' lambda statistic of 0.031, with a p-value of 0.000.

TABLE 9
UNSTANDARDIZED COEFFICIENTS OF THE CANONICAL DISCRIMINANT FUNCTION (TWO YEARS BEFORE)

	Function 1
EBIT.TA	0.5951
LnTA	0.5051
L.TD	-0.0237
S.CAP	-0.002
S.CA	0.00016
NACE_1	0.4388
NACE_3	0.4918
NACE_4	0.4951
LIMITED CORP.	0.5951
(Constant)	-4.0845

Table 10 shows the confusion matrices and errors for both training and test data. In this case, the error percentages are 18.539% and 22.881%, respectively. These errors are slightly higher than the errors obtained by the model *one year before* failure. Consequently, the model has an accuracy of 77.12% when it comes to estimating the situation of new firms, that is to say, predicting whether the firm will fail in the next two years.

TABLE 10
CONFUSION MATRICES AND ERRORS WITH LDA (TWO YEARS BEFORE)

PREDICTED CLASS	*ACTUAL CLASS*				
	LDA (two years before)	Training	18.538%	Test	22.881%
		Failed	*Healthy*	*Failed*	*Healthy*
	Failed	406	109	95	31
	Healthy	66	363	23	87

Having obtained the results from the application of the LDA model, we compare them to the AdaBoost results for the same period. As in the first case, we construct the AdaBoost classifier with one hundred trees grown to a maximum depth of 2. The error in the test data set is 10.169%, which implies a reduction of 55% in comparison to the LDA error. Moreover, as we noted in the *one year before* model, most errors occurred when classifying healthy firms as failed. In other words, with the AdaBoost classifier the Type I error is lower than Type II. Table 11 shows the confusion matrices and error percentages for the two data sets.

TABLE 11
CONFUSION MATRICES AND ERRORS WITH ADABOOST (TWO YEARS BEFORE)

	ACTUAL CLASS				
PREDICTED CLASS	*AdaBoost* (two years before)	Training	9.640%	Test	22.881%
		Failed	*Healthy*	*Failed*	*Healthy*
	Failed	462	81	114	20
	Healthy	10	391	4	98

Once again we check the significance of the differences between the LDA and the AdaBoost models through 50 experiments (five repetitions of 10-fold cross-validation) using the 1180 firms in the entire sample. Table 12 shows the normality tests as well as the mean and standard deviation of the error across the 50 experiments and for both classifiers. After applying the test for the difference of means between two normal populations the result is very clear (t-statistic of -13.6623 and its p-value 0.0000). By way of conclusion, we can once again consider the differences to be statistically significant with the AdaBoost model ahead in the year and two years before failure.

TABLE 12
NORMALITY TESTS, AVERAGE AND STANDARD DEVIATION OF ERROR (TWO YEARS BEFORE)

Test	*LDA*	*AdaBoost*
Kolmogorov-Smirnov	0.1355	0.1182
KS. p value	0.3176	0.4871
Shapiro	0.9489	0.9764
Shapiro p value	0.0308	0.4120
Average error	0.1946	0.1086
Standard deviation	0.0346	0.0280

Using the aforementioned function, we can study the relative importance of the variables. In this case, the most important ones are CF.TD, C.CL, EBIT.TA and lnTA with values of the relative important measure of 24.59, 18.54, 14.11 and 12.9%, respectively. It is worth mentioning that in this case the number of splits is more concentrated than in the year before model and therefore six ratios participate in less than 1% of the splits. Table 13 shows the ranking of variables ordered from highest to lowest relative importance. The most important variables for the AdaBoost classifier are the capacity to return liabilities, the acid test, economic profitability and the proxy variable of the size of the firm. These two last variables were also found among the most significant variables for the *one year before* model. The capacity to return liabilities represents the amount of available liquidity (resulting from the income during the business year once all demandable expenses necessary for obtaining such an income have been paid) for paying back loans for the total debt and capital; the higher the index, the higher the capacity to repay financing which means that the firm is less likely to fail. The acid test, for its part, indicates the firm's ability to keep up with the required payments of current liabilities deriving from the productive and commercial cycle. A high ratio would mean an excess of cash, which is detrimental to efficiency and final profitability, although it would depend on the firm's activity and its current state. On the other hand, a very low index value, much less than one, would imply inability to keep up with payment obligations. By way of conclusion, an intermediate value would be desirable, neither too high nor too low.

TABLE 13
RELATIVE IMPORTANCE OF VARIABLES (TWO YEARS BEFORE)

Variable	*Relative importance*	*Variable*	*Relative importance*
CF.TD	24.59	C.TA	2.82
C.CL	18.55	WC.TA	1.21
EBIT.TA	14.11	CA.CL	0.81
LnTA	12.90	S.TA	0.81
S.CAP	8.47	S.CA	0.81
NACE	6.45	LEGAL STR.	0.40
L.TD	4.44	WC.S	0.40
EBIT.CAP	3.23	CA.TA	0.00

5. CONCLUSIONS

The task of predicting corporate failure is still valid forty years after Beaver and Altman's pioneering models. In this paper, we have proposed an alternative and novel machine learning model that can be applied to failure prediction. More specifically, we have analyzed the AdaBoost ensemble classifier method which shows an important improvement in terms of both theoretical and empirical accuracy. As previously explained, AdaBoost is based on an iterative construction of classifiers on modified versions of the training data set which are generated according to the error of the classifier trained in the previous iteration and focusing on the hardest examples of the training data set.

For empirical application, we have worked with the two classes (as is usual in failure prediction publications) and we therefore distinguish between healthy and failed firms. We have used the concept of legal failure to include those firms which had gone into receivership or bankrupt during the period 2000-2003 whenever they fulfilled the requirements described in Section 3 of this article. Traditional linear discriminant analysis obtains errors of 20.34% and 22.88% in the cases of one and two years before failure, respectively. While these results are not bad, those obtained by the novel classifier AdaBoost are far better, with admirable errors of 8.898% and 10.17% for the test sets and both periods. The gain is evaluated at around 55% against the LDA model.

We have also proved that these differences in favour of AdaBoost are statistically significant in both periods by conducting 50 experiments and using the corresponding test to compare means. The main conclusion that we can draw from the analysis is of the great potential of the proposed technique which, while at first appearing difficult to apply, we have shown to be as easy to implement with the appropriate software tools as other statistical techniques. It is also worth mentioning that the validation of results for the two years before model allows us to forecast failure further in advance than with the *one year before* model.

On the other hand, thanks to the relative important measure, which is one of our contributions, a better interpretation of the model can be achieved. We believe that this measure can play a very important role in the popularization of the method. In this way, it is possible to know which variables exert the most influence on the model and we can rank the variables from greatest to least importance. In this paper, the most influential variables are economic profitability and the logarithm of total assets as indicators of the size of the firm, variables common to both periods. In addition to these two variables, in the *one year before* model it is also worth mentioning the level of indebtedness and efficiency whereas in the two years before model we must include the capacity to return liabilities and the acid test.

In this research, we have compared the AdaBoost algorithm with the Linear Discriminant Analysis method because this is the most popular published technique for corporation failure prediction. Nevertheless, in future lines of research, we will compare it with other well-known classification methods such as logistic regression or artificial neural networks. We would also like to explore the addition of further years to the time horizon since this work is limited to one and two years before failure.

REFERENCES

ABAD, C.; ARQUERO, J. L., and JIMÉNEZ, S. 2004. Procesos de fracaso empresarial. Identificación y contrastación empírica, *La Contabilidad sin Fronteras,* ASEPUC, Granada.

ALFARO, E.; GÁMEZ, M., and GARCÍA RUBIO, N. 2006. Adabag: Applies AdaBoost.M1 and Bagging, *R package,* version 1.0.
http://cran.r-project.org/src/contrib/Descriptions/adabag.html (accesed 9 November 2007).

ALTMAN, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance* 23 (4), 589-609.

BAUER, E., and KOHAVI, R. 1999. An empirical comparison of voting classification algorithm: Bagging, boosting and variants, *Machine Learning* 36, 105-142.

BEAVER, W. H. 1966. Financial ratios as predictors of failure, *Empirical Research in Accounting: Selected Studies*, Supplement to Volume 5 of *Journal of Accounting Research,* 71-111.

BREIMAN, L. 1998. Arcing classifiers. *The Annals of Statistics* 26(3), 801-849.

BRYAN, M.; TIRAS, L., and WHEATLEY, M. 2002. The interaction of solvency with liquidity and its association with bankruptcy Emergence, *Journal of Business Finance & Accounting* 29, 935-965.

CHARALAMBOUS, C.; CHARITOU, A., and KAOROU, Y. F. 2000. Comparative analysis of artificial neural network models: Application in bankruptcy prediction, *Annals of Operation Research* 99(4), 403-425.

DIETTERICH, T. G. 2000. Ensemble methods in machine learning, In *Multiple Classifier Systems,* Cagliari, Italy.

DRUCKER, H., and CORTES, C. 1996. Boosting decision trees, In TOURETZKY, D. S.; MOZER, M. C., and HASSELMO, M. E., Editors, *Advances in Neural Information Processing Systems 8: Proc. of NIPS'95,* 8, 479-485. The MIT Press.

FERRANDO, M., and BLANCO, F. 1998. La previsión del fracaso empresarial en la comunidad valenciana: una aplicación de los modelos discriminante y logit, *Revista Española de Financiación y Contabilidad* 95 (4), 499-540.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems, *Annals of Eugenics* 7, 179-188.

Freund, Y., and Schapire, R. E. 1996. Experiments with a new boosting algorithm. In Proc. 13th International Conference on Machine Learning, 148-146. Morgan Kaufmann.

Freund, Y., and 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.

— 1998. Discussion of the paper arcing classifiers of Leo Breiman, *The Annals of Statistics* 26(3), 824-832.

Friedman, J.; Hastie, T., and Tibshirani, R. 2000. Additive logistic regression: a statistical view of boosting, *Annals of Statistics* 38(2), 337-407.

Frydman, H.; Altman, E., and Kao, D. 1985. Introducing recursive partitioning for financial classification: the case of financial distress, *Journal of Finance*, 269-291.

Gabás, F. 1990. Técnicas Actuales de Análisis Contable. Evaluación de la Solvencia Empresarial, Madrid: Instituto de Contabilidad. y Auditoría de Cuentas. Ministerio de Economía y Hacienda.

Gallego, A. M., y Gómez, A. 2002. Análisis integrado de la absorción y quiebra empresarial mediante la estimación de un modelo multilogit, *Revista Española de Financiación y Contabilidad* 31 (1), 111-144.

Hand, D. J. 1981. *Discrimination and classification,* Ed. John Wiley.

Labatut, G.; Molina, R., and Pozuelo, J. 2006. La predicción del fracaso empresarial en las PYMEs: un factor clave en el nuevo contexto financiero internacional, *VI Jornadas sobre Predicción de la Insolvencia Empresarial,* AECA, Carmona (Sevilla).

Manzaneque, M. 2006. Modelos de pronóstico de la insolvencia empresarial, *Tesis Doctoral,* Universidad de Castilla-La Mancha.

Olshon, J. A. 1980. Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research* 18, (1), 5-12.

Pérez, M. 2006. Artificial neural networks and bankruptcy forecasting: a state of the art., *Neural Computing & Applications* 15(2), 154-163.

Ravi Kumar, P., and Ravi, V. 2007. Bankruptcy prediction in banks and firms via statistical and intelligent techniques-A review, *European Journal of Operational Research,* 180(1), 1-28.

Rodríguez, M. 2001a. Predicción del fracaso empresarial en compañías no financieras. Consideración de técnicas de análisis multivariante de corte paramétrico, *Actualidad Financiera,* junio, 27-42.

Rodríguez, M. 2001b. El fracaso empresarial en Galicia a través del análisis tradicional de la información contable. La capacidad predictiva del análisis univariante, *II Congreso de Economía de Galicia*, Santiago de Compostela.

Schapire, R. E. 2002. The boosting approach to machine learning: an overview, *En Workshop on Nonlinear Estimation and Classification*, MSRI.

Sanchís, A.; Gil, J. A., and Heras, A. 2003. El análisis discriminante en la previsión de la insolvencia en las empresas de seguros no vida, *Revista Española de Financiación y Contabilidad* 116, enero-marzo, 183-233.

Serrano, C. 1997. Feedforward neural networks in the classification of financial information, *European Journal of Finance* 3(3), 183-202.

Uriel, E. 1995. *Análisis de datos: series temporales y análisis multivariante,* Editorial AC.

Wilson, R. L., and Sharda, R. 1994. Bankruptcy prediction using neural network, *Decision Support Systems* 11, 545-557.

Zmijewski, M. 1984. Methodological issues related to the estimation of financial distress prediction models, *Journal of Accounting Research,* 22, 59-86.

APPENDIX 1

TABLE 14
TEMPORAL DISTRIBUTION OF FAILURES

	2000	2001	2002	2003	Total
Receivership	40	42	54	49	185
Bankruptcy	94	80	114	117	405

APPENDIX 2

a) *Profitability Ratios*

Earnings before interest and taxes/Total Assets (EBIT.TA).
Earnings before taxes/Capital (EBIT.CAP).

b) *Economic structure ratios*

Current assets/ Total assets (CA.TA).
Cash/Total Assets (C.TA).

c) *Financial structure ratios*

Sales/Capital (S.CAP).

d) *Solvency, indebtedness and liquidity ratios*

Current Assets/Current Liabilities (CA.CL).
Cash/Current liabilities (C.CL).
Working Capital/Sales (WC.S).
Liabilities/Total Debt (L.TD).
Cash Flow/Total Debt (CF.TD).
Working Capital/ Total Assets (WC.TA).

e) *Eficciency and productiviry ratios*

Sales /Total Assets (S/TA).
Sales /Current Assets (S/CA).

f) *Other ratios*

Proxy variable of the size of the firm. Logarithm of Total Assets (ln TA).
NACE code at level 1 (NACE1).
Firm's legal structure (LEGAL STRUCTURE).

APPENDIX 3

TABLE 15
COMPOSITION OF THE SAMPLE BY SECTOR

NACE1		0	1	2	3	4	5	6	7	8	9
Healthy	N	13	76	124	28	53	236	14	38	3	5
	%	72.22	40.21	47.33	45.16	38.41	58.13	50.00	61.29	42.86	62.50
Failed	N	5	113	138	34	85	170	14	24	4	3
	%	27.78	59.79	52.67	54.84	61.59	41.87	50.00	38.71	57.14	37.50
Total	N	18	189	262	62	138	406	28	62	7	8
	%	1.53	16.02	22.20	5.25	11.69	34.41	2.37	5.25	0.59	0.68

Table 16
NACE codes at level 1

Digit	Code	Title
0	A	Agriculture, hunting and forestry
0	B	Fishing
1	C	Mining and quarrying
1	DA	Manufacture of food products, beverages and tobacco
1	DB	Manufacture of textiles and textile products
1	DC	Manufacture of leather and leather products
2	DD	Manufacture of wood and wood products
2	DE	Manufacture of pulp, paper and paper products; publishing and printing
2	DF	Manufacture of coke, refined petroleum products and nuclear fuel
2	DG	Manufacture of chemicals, chemical products and man-made fibres
2	DH	Manufacture of rubber and plastic products
2	DI	Manufacture of other non-metallic mineral products
2	DJ	Manufacture of basic metals and fabricated metal products
2	DK	Manufacture of machinery and equipment n.e.c.
3	DL	Manufacture of electrical and optical equipment
3	DM	Manufacture of transport equipment
3	DN	Manufacturing n.e.c.
4	E	Electricity, gas and water supply
4	F	Construction
5	G	Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods
5	H	Hotels and restaurants
6	I	Transport, storage and communication
6	J	Financial intermediation
7	K	Real estate, renting and business activities
7	L	Public administration and defence; compulsory social security
8	M	Education
8	N	Health and social work
9	O	Other community, social and personal service activities
9	P	Activities of households
9	Q	Extra-territorial organizations and bodies

Table 17
Sample composition by legal structure

Legal structure	Total		Healthy		Failed	
	N	%	N	%	N	%
Public corp.	666	56.44	318	47.75	348	52.25
Limited corp.	501	42.46	262	52.30	239	47.70
Other corp.	13	1.10	10	76.92	3	23.08

Table 18

Sample composition by firm size. Two groups have been created: big firms, those with a logarithm of total assets above the mean, and small firms, those with a logarithm of total assets below the mean. This criterion has been previously applied by Manzaneque (2006) and Bryan *et al.* (2002)

Size	*Total*		*Healthy*		*Failed*	
	N	%	N	%	N	%
Big	548	46.44	275	50.18	273	49.82
Small	632	53.56	315	49.84	317	50.16