Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 23 (2015) 9 – 14

Procedia Economics and Finance

www.elsevier.com/locate/procedia

2nd GLOBAL CONFERENCE on BUSINESS, ECONOMICS, MANAGEMENT and TOURISM, 30-31 October 2014, Prague, Czech Republic

A Comparison of Classification/Regression Trees and Logistic Regression in Failure Models

Irimia-Dieguez, A.I.[a]*, Blanco-Oliver, A.[a], Vazquez-Cueto, M.J. [a]

[a]*Faculty of Economics and Business Administration, University of Seville, Av. Ramón y Cajal, s/n, 41018Seville (Spain)*

Abstract

The use of non-parametric statistical methods, the development of models geared towards the homogeneous characteristics of corporate sub-populations, and the introduction of non-financial variables, are three main issues analysed in this paper. This study compares the predictive performance of a non-parametric methodology, namelyClassification/Regression Trees (CART), against traditional logistic regression (LR) by employing a vast set of matched-pair accounts of the smallest enterprises, known as micro-entities,from the United Kingdom for the period 1999 to 2008 that includes financial, non-financial, and macroeconomic variables. Our findings show that CART outperforms the standard approach in the literature, LR.

Selection and/ peer-review under responsibility of Academic World Research and Education Center

*Keywords:*Micro-entities, Failure models, Financial ratios, Non-financial information, Macroeconomic variables, Logistic regression, Classification/Regression Trees.

1. Introduction

In recent years, three new research lines have appeared which strive to improve the performance of bankruptcy models: (a) introducing non-financial information as predictor variables (Grunert, Norden & Weber, (2005); (b) developing models specifically designed for each firm feature, such as size and sector (Altman, &Sabato, 2007); and (c) implementing non-parametric statistical techniques to fit the bankruptcy models (Jagric, Kracun &Jagric, 2011).

This study deals with these three advances developing a failure model specifically designed for the smallest micro-enterprises: micro-entities (hereafter, MEs), which have recently been defined by the Competitiveness Council of the European Union as those companies with an annual turnover of less than €700,000, total assets

*Irimia-Dieguez, A.I..Tel.: +34-954-559875; fax: +34-954-557570.
E-mail address:anairimia@us.es

Selection and/ peer-review under responsibility of Academic World Research and Education Center
doi:10.1016/S2212-5671(15)00493-1

of less than €350,000, and average number of employees during the financial year of no more than 10, (Official Journal of the European Union, 2012).The particular characteristics of MEs: (a) higher probability of failure (Carter, & Van Auken, 2006); (b) great limitation of publicly available financial information, due to the fact that they file abridged accounts; and (c) the inexistence of failure models adapted to this types of firms despite of their leading role in the economic activity worldwide, justify the need of the development of specific failure models. Furthermore, we also test the accuracy capacity to detect the failure of a non-parametric statistical technique Classification/Regression Trees (CART), in comparison with the classic logistic regression (LR) analysis. In this sense, previous literature shows that CART often outperforms LR in the failure environment .

After the implementation of the Basel Accord regulation (Basel II), considerable studies have been undertaken in an effort to predict the failure of SMEs. Whereas the importance of financial factors is widely accepted, the relevance of non-financial predictors appears to need more empirical evidence. Moreover, nothing is known about the applicability of default prediction models to MEs, and whether non-financial information improves the predictive capacity of models developed specifically for them due to the lack of research that deals with these kinds of firms.

One of the most relevant models specifically made for SMEswas developed by (Altman, & Sabato, 2007). Their study compares the traditional Z-score model with two new models which consider other financial variables and use traditional logistic regression. On a panel of data of over 2,000 US SMEs in the period 1994-2002, these authors find that the new models outperform the traditional Z-score model by almost 30 per cent, in terms of prediction power. Based on the above research, (Altman, &Sabato, 2007)they explore the effect of the introduction of non-financial information as predictor variables into the models developed by (Altman, & Sabato, 2007). Employing a large sample (5.8 million) of sets of accounts of unlisted firms from the U.K. in the period 2000-2007, they find that non-financial information makes a large contribution (by approximately 13% in terms of the area under the receiver operating characteristics curve, henceforth, AUC) towards increasing the default prediction power of risk models.

Therefore, the main objective of this study is to compare LR and CART in the building of a failure model designed for MEs which introduce financial, non-financial and macroeconomic variables. The large size of the sample (almost 40,000 set of accounts of MEs) is an important strength for the reliability of our findings. Moreover, the use of a parsimonious model constitutes a noteworthy improvement.

In Section 2, we provide details of our sample and methodology carried out. In Section 3, several failure models for MEs are developed, comparing LR and CART approaches. In Section 4, the results are shown and discussed them. Finally, Section 5 provides the main conclusions and future lines of research.

2. Data set

A dataset provided by a U.K. Credit Agency is used in this study. After eliminating missing and abnormal cases and selecting a random sample of MEs, 39,710 sets of accounts of MEs (50% non-failed) for the period 1999-2008 remained.In line with other studies, we define corporate failure as entry into liquidation, administration or receivership between 1999 and 2008. The accounts analyzed for failed companies are the last set of accounts filed in the year preceding insolvency.For each case, the dependent variable takes the value 1 when the ME failed, and 0 otherwise.Finally, to run the models, our final dataset was randomly split into three sub-sets; a training set of 60%, a validation set of 20%, and a test data set (or hold-out sample) of 20% (Hastie, Tibshirani & Friedman, 2009).

Table 1 describes the variables considered in this study and the theoretical relationship with the failure of the firm. All the financial ratios used in this study were employed in prior research, such as Altman, (1968); Altman, Sabato & Wilson, (2010); Ohlson, (1980); Taffler, (1984) and Zmijewski (1984). Moreover, based in the findings of Carter, Van Auken, (2006), it seems reasonable to assume that an adequate failure model made specifically for MEs should also introduce non-financial information. Finally, since several studies have shown a positive relationship between the adverse economic cycle and the number of corporate failures Moon, and Sohn (2010), we also include a macroeconomic variable (*Industry_solvency*) which measure the financial health of the sector in which operate the firm and is inverse of the probability of bankruptcy of the sector.

Table 1.Financial, non-financial and macroeconomic variables

Variable	Abbreviation	Category	Theoretical relationship bankruptcy
Financial Ratios			
Capital employed / Total liabilities	Celt	Leverage	-
Short-term liabilities / Total assets	Stlta	Leverage	+
Total liabilities / Current assets	Tlca	Leverage	+
Net worth / Total assets	Nwta	Leverage	-
Quick assets / Current assets	Qaca	Liquidity	-
Cash / Net worth	Cashnt	Liquidity	-
Current assets / Current liabilities	Cacl	Liquidity	-
Cash / Total assets	Cashta	Liquidity	-
Retained profit / Total assets	Rpta	Profitability	-
Trade creditors / Trade debtors	Tctd	Activity	+
Trade creditors / Total liabilities	Tctl	Activity	+
Trade debtors / Total assets	Tdta	Activity	+
Nepierian logarithm total assets	Ln_asset	Size	+/-
Total assets	T_asset	Size	+/-
Non-financial and Macroeconomic Variables			
Audited accounts	Audited	No (0)	+
		Yes (1)	-
Positive judgment audit report	Aq_clean	No (0)	+
		Yes (1)	-
Negative judgment audit report	Aq_no_clean	No (0)	-
		Yes (1)	+
Change auditor	Change_auditor	No (0)	-
		Yes (1)	+
Number of legal claims	Number_LCs		+
Value of legal claims	Value_LCs		+
Late filing days	Late_filing_day		+
Napierian logarithm age	Ln_age		-
Charge on assets	Charge_asset	No (0)	-
		Yes (1)	+
Family firm	Family_firm	No (0)	-
		Yes (1)	+
Industry solvency	Industry_solvency		-

3. Methodology

3.1.Logistic regression (LR)

To obtain a parsimonious failure model, the accuracy ratio (AR) is observed for each financial variable. To avoid the problem of multicollinearity between the independent variables of the model, (Altman, &Sabato, 2007) suggest thatonly one variable is selected from each ratio category. The most significant financial ratios were *Celt, Cashta, Rpta, Tdta and Ln_Asset.*To select the most significant non-financial and macroeconomic variables a forward stepwise selection procedure was implemented, thereby concluding that *Number ccjs, Late_Filing_Days, Ln_Age and Family Firm* are the most significant non-financial variables. The variable *Industry_Solvency* was also significant (p-value <0.05). The coefficients and significance level of all the variables finally considered in our model are collected in Table 2. As shown in this table, all the slopes (signs) follow our expectations. The relevance of these variables on the failure of firms can also be analysed by means of the absolute values of the Wald ratio coefficients of each variable.*Cashta* and *Ln_Asset* are the most relevant variables in the model which consider only financial predictors (Model LR1). The variables with the greatest predictive power of Model LR2 are *Number_ccjs* and *Late_Filing_Days.*

Table 2.Logistic-default prediction models for the micro-entities

Abbreviation	Variable	Categ.	Logistic Regression Model 1 (LR 1) Coef.	Wald	Sig.	Logistic Regression Model 2 (LR 2) Coef.	Wald	Sig.
Celt	Capital employed / Total liabilities	F	-0.054	179.92	0.000	-0.031	59.421	0.000
Cashta	Cash / Total assets	F	-1.929	1477.66	0.000	-1.504	781.36	0.000
Rpta	Retained profit / Total assets	F	-0.385	834.93	0.000	-0.374	771.62	0.000
Tdta	Trade debtors / Total assets	F	0.420	94.90	0.000	0.551	144.06	0.000
Ln_asset	Ln total assets	F	0.804	1317.83	0.000	0.808	1175.40	0.000
Number_LCs	Number of legal claims	NF				1.681	695.22	0.000
Late_filing_day	Late filing days	NF				0.006	439.35	0.000
Ln_age	Ln age	NF				-0.298	242.91	0.000
Family_firm	Family firm	NF				0.266	98.56	0.000
Indwoe	Industry solvency	ME				-0.626	508.48	0.000
	Intercept		-7.955	1183.33	0.000	-6.298	538.04	0.000

F=Financial; NF=Non-Financial;ME=Macro-economic

3.2. Classification/Regression Trees (CART)

From the application of a CART algorithm, a classification tree was built with an initial node composed of 23,144 firms and with only those ten variables used in Model LR2. By using the Gini impurity function, the prior probabilities observed in the sample, equal cost of misclassification for both groups, and the 0SERULE rule, we obtained twelve trees with their associated validation and replacement costs. The best tree is that with 28 nodes, and validation and replacement costs of 0.54868 (+/- 0.00587) and 0.47748, respectively. This tree (see appendix 1 for a detailed description) decreases the percentage of incorrect classification in the training sample, and obtains suitable performance in the validation sample. Moreover, this model also offers a clear interpretation of the results despite the reduced number of nodes.

For this tree, a test accuracy (TA) of 76.18%, and type I-II errors of 24.67% and 22.97% respectively, are obtained in the training sample; the AUC is 0.816. In the test sample, the TA is of 72.63%, the type I-II errors are of 26.65% and 28.09% respectively, and the AUC is equal to 0.771. In addition, the software determines the relative importance of each variable within the construction of the tree, which is given as *Rpta*(100.00%), *Celt*(94.14%), *Cashta*(79.64%), *Late_Filing_Days*(47.68%), *Number_ccjs*(38.48%), *Industry_Solvency*(31.65%), *Tdta*(27.96%), *Ln_Asset*(16.20%), *Ln_Age*(1.31%), and *Family_Firm*(0.85%).

The three most important variables are financial ratios, followed by three non-financial variables, with almost half of the importance. Whether the ME is a family firm remains irrelevant in the tree construction.

4. Results and Discussion

To evaluate the performance of the failure models developed here, the area under the ROC curve (AUC) is used. Furthermore, in accordance with West, (2000), the expected misclassification cost (EMC) is also employed as performance criteria. Table 3 below contains the AUC, Type I-II errors and EMC of all the models built. the values selected for the calculation of the misclassification costs are: C_{21}=1 and C_{12}=5 (as recommended by West, (2000); P_{21} and P_{12} are dependent of each model; and $\hat{\pi}_1 = 0.482$ and $\hat{\pi}_2 = 0.518$.

As can be observed in this table, when the non-financial and macroeconomic variables are considered to predict the failure (LR2), all the performance criteria are clearly improved respect to consider only financial ratios as input variables (LR1). Nevertheless, the prediction power of LR2 is clearly increased by using CART. Exactly, our findings reveal that CART detects better the failure of a firm than LR, with a improvements in terms of AUC and EMC by 0.7% (-3.1%) and 20.86% (7.49%) in training sample (test sample). Large differences are observed in terms of the EMC criteria, and therefore, CART approach is a way to reduce the number of incorrect decisions on

failure firms (and then decrease the monetary losses). Thus, we conclude, in line with other authors (e.g. Gepp, Kuldeep, & Bhattacharya, 2009), that, in general, not only does CART method has a greater AUC, but it also incurs in lower EMC than the traditional LR approach.

Based in the previous empirical evidences, we suggest both the inclusion of non-financial and macroeconomic variables and the implementation of CART (instead of widely employed LR) to detect the failure of the smallest business.

Table 3. AUC, TA, Type I errors and Type II errors

		Model		
		LR 1	LR 2	CART
Training sample	AUC	0.736	0.809	0,816
	Test accuracy	70.22%	74.08%	76.18%
	Type I	31.49%	24.54%	24.67%
	Type II	29.05%	29.54%	22.97%
	EMC	0.8857	0.8634	0.6548
Test sample	AUC	0.770	0.806	0.775
	Test accuracy	70.74%	72.99%	73.13%
	Type I	30.97%	24.83%	26.05%
	Type II	27.77%	28.69%	27.68%
	EMC	0.8513	0.8438	0.7689

5. Conclusions

This study develops a failure model for MEs by using only their publically available financial ratios for this firm-size and also including non-financial and macroeconomic information. Our results show, firstly, that the combined use of financial, non-financial and macroeconomic variables improves the capacity of our model to predict the bankruptcy of MEs. Secondly, the findings also confirm the theoretical superiority of non-parametric statistical techniques (CART) on the classic logistic regression (LR) analysis. The empirical evidence shows that the CART clearly improve the AUC and EMC by 0.7% (-3.1%) and 20.86% (7.49%) in training sample (test sample).In addition, CART models have the advantage of its transparency.

Appendix

Description of the Classification Tree.

NODE	SPLIT ON	DESCRIPTION
Node 1	CASHTA	A case goes left (Node 2) if CASHTA <= 0.29527; otherwise goes Node 20
Node 2	CETL	A case goes left (Node 3) if CETL <= 0.49074; otherwise goes Node 18
Node 3	NO_CCJS	A case goes left (Node 4) if NO_CCJS = 0; otherwise goes to **Terminal Node 16 (FAILED)**
Node 4	LN_ASSET	A case goes left (Node 5) if LN_ASSET <= 10.06453; otherwise goes Node 14
Node 5	LAST_ACC_LATE	A case goes left (Node 6) if LAST_ACC_LATE = 0; otherwise goes right to **Terminal Node 10 (FAILED)**
Node 6	CETL	A case goes left (Node 7) if CETL <= -0.09233; otherwise goes Node 10
Node 7	LAST_ACC_LATE	A case goes left to **Terminal Node 1 (NON-FAILED)** if LAST_ACC_LATE#0; otherwise goes Node 8
Node 8	TDTA	A case goes left (Node 9) if TDTA <= 0.13485; otherwise goes to **Terminal Node 4 (FAILED)**
Node 9	INDWOE	A case goes left to **Terminal Node 2 (NON-FAILED)** if INDWOE <= -0.38632; otherwise goes to **Terminal Node 3 (NON-FAILED)**
Node 10	LAST_ACC_LATE	A case goes left (Node 11) if LAST_ACC_LATE =0; otherwise goes to **Terminal Node 9 (FAILED)**
Node 11	AGERISK_LOGIT	A case goes left to **Terminal Node 5 (NON- FAILED)** if AGERISK_LOGIT = (2, 3); otherwise goes Node 12
Node 12	FAMILY	A case goes left to **Terminal Node 6 (NON-FAILED)** if FAMILY = 1; otherwise goes Node 13
Node 13	TDTA	A case goes left to **Terminal Node 7 (FAILED)** if TDTA <= 0.95119; otherwise goes to **Terminal Node 8 (NON-FAILED)**
Node 14	LAST_ACC_LATE	A case goes left (Node 15) if LAST_ACC_LATE=0; otherwise goes to **Terminal Node 15 (FAILED)**
Node 15	TDTA	A case goes left (Node 16) if TDTA <= 0.07768; otherwise goes to **Terminal Node 14 (FAILED)**
Node 16	LAST_ACC_LATE	A case goes left (Node 17)if LAST_ACC_LATE=0; otherwise goes to **Terminal Node 13 (FAILED)**

Node 17	RATIO_PRTA	A case goes left to **Terminal Node 11 (Clase 1)** if RATIO_PRTA <= -0.27889; otherwise goes to **Terminal Node 12 (NON- FAILED)**
Node 18	LAST_ACC_LATE	A case goes left (Node 19) if LAST_ACC_LATE=0; otherwise goes to **Terminal Node 19 (FAILED)**
Node 19	NO_CCJS	A case goes left to **Terminal Node 17 (Clase 0)** if NO_CCJS = 0; otherwise goes to **Terminal Node 18 (FAILED)**
Node 20	NO_CCJS	A case goes left (Node 21) if NO_CCJS = 0; otherwise goes to **Terminal Node 28 (FAILED)**
Node 21	PRTA	A case goes left (Node 22) if PRTA <= 0.03252; otherwise goes Node 27
Node 22	LAST_ACC_LATE	A case goes left (Node 23) if LAST_ACC_LATE=0; otherwise goes to **Terminal Node 25 (FAILED)**
Node 23	LN_ASSET	A case goes left to **Terminal Node 20 (NON- FAILED)**, if LN_ASSET <= 9.82516; otherwise goes Node 24
Node 24	CASHTA	A case goes left (Node 25) if CASHTA <= 0.71807; otherwise goes to **Terminal Node 24 (NON-FAILED)**
Node 25	INDWOE	A case goes left to **Terminal Node 21 (FAILED)** if INDWOE <= 0.12; otherwise goes Node 26
Node 26	PRTA	A case goes left to **Terminal Node 22 (FAILED)** if PRTA <= -0.20; otherwise goes to **Terminal Node 23 (NON- FAILED)**
Node 27	LAST_ACC_LATE	A case goes left to **Terminal Node 26 (NON-FAILED)** if LAST_ACC_LATE=0; otherwise goes to **Terminal Node 27 (FAILED)**

References

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy.*Journal of Finance*, *23* (4), 589-609.

Altman, E. I., Sabato, G., & Wilson, N. (2010). The Value of Non-financial Information in Small and Medium-sized Enterprise Risk Management.*Journal of Credit Risk*,*6* (2), 95-127.

Altman, E.I., &Sabato, G. (2007). Modeling Credit Risk for SMEs: Evidence from the U.S. Market.*Abacus,* 43 (3), 332-357.

Carter, R. & Van Auken, H. (2006). Small Firm Bankruptcy.*Journal of Small Business Management, 44 (4), pp. 493-512.*

Carter, R.; Van Auken, H. (2006). Small Firm Bankruptcy.*Journal of Small Business Management, vol. 44*(4), 493–512.

Gepp, A.,Kuldeep, K., & Bhattacharya, S. (2009). Business failure prediction using decision trees.*Journal of Forecasting, 6, 536-555.*

Grunert, J., Norden, L.& Weber, M. (2005). The Role of Non-Financial Factors in Internal Credit Ratings.*Journal of Banking and Finance*, *29* (2), 509-531.

Hastie, T.; Tibshirani, R.; Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. New York.

Jagric, V.,Kracun, D., &Jagric, T. (2011).Does Non-linearity Matter in Retail Credit Risk Modeling? *Czech Journal of Economics and Finance, 61*(4), 384-402.

Moon, T., and S. Sohn (2010). 'Technology Credit Scoring Model Considering both SME Characteristics and Economic Conditions: The Korean Case,' *Journal of the Operational Research Society*, *61* (4), 666-675.

Official Journal of the European Union (2012). Directive 2012/6/EU of the European Parliament and of the Council.http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:081:0003:0006:EN:PDF. Accessed on Jan. 10, 2013

Ohlson, J.A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy.*Journal of Accounting Research*, *18* (1), 109-131.

Taffler, R. J. (1984). Empirical Models for the Monitoring of U.K. Corporations.*Journal of Banking and Finance*, *8* (2), 199-227.

West, D. (2000). Neural Network Credit Scoring Models.*Computers and Operations Research, 27* (11-12), 1131-1152.

Zmijewski ME (1984). Methodological Issues related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research,* 22(Suppl.):59-82.