Bankruptcy prediction in Colombian case, using multilayer perceptron trained with memetic algorithm

Student:Iván Andrés Trujillo Abella,
Proffesors: Eliana María González Neira & Gabriel Mauricio Zambrano Rey

***Abstract*—Literature about Bankruptcy prediction is still incipient, therefore this work try to fill this gap by using machine learning and metaheuristics techniques to find an optimal set of weights in a MLP model.**

***Index Terms*—Machine learning, Bankruptcy, Metaheuristics, Evolutionary Algorithms, Local Search, Memetic algorithms, Neural Networks, Multilayer Perceptron.**

I. Information(Data)

Data was retrieved from SIS by the period 2016-2019, classifying as bankrupt all Small and Medium Enterprises(SMEs) firms that were declared some process different to active or preoperative state.

To predict bankruptcy, financial ratios from one and two years before the occurrence of the event were used(in different models). For instance, for those firms that entered into a process in 2019, the financial ratios of 2018 and 2017 were used to train the models.

For those firms that did not experience the event (controls), the mean of the financial ratios was used based on their period of activity. For instance, if firm j entered into database in 2017 and did not experience the event, the mean of the financial ratios for the period (2017 - 2019) was used.

A. Exclusion criteria

- Financial information reported on a date different from December 31 of each year was excluded.
- Firms declared to be in a preoperative condition each year were excluded from the database.
- Firms that presented the event for the initial year(2016) were excluded from the database.
- Firms that present missing values and not valid financial ratios were excluded from database.

II. Description of data

To describe categorical data, absolute and relative frequencies were used. The numerical variables were described using the mean and standard deviation if the variable is normally distributed; otherwise, the median, 25th percentile, and 75th percentile were used.

To compare the financial ratios among non-bankrupt and bankrupt firms and determine if there are significant differences among them, we used t-tests or the Wilcoxon rank sum test if the variable is normal or not, respectively. To test the normality of variables, we used the Shapiro-Wilk test. To determine if there is independence between the event and categorical variables, the χ^2 test was used.

Finally, to test if there is correlation(monotonicity) among numerical variables the spearman coefficient was used.

A. Variables (Financial ratios)

Financial ratios could be categorized in some groups, for instance *liquidity, leverage* and *profitability* for this work were used:

1) liquidity ratios: Measure the firm ability to pay its financial obligations; Current ratio and Ratio of short-term liabilities were used.

2) Profitability ratios: Measure the firm ability to generate incomes; Gross profit margin, Net profit margin, Return on equity and Return on asset were used.

3) Leverage ratios: Measure the level of debt in long run; Debt equity ratio and Indebtedness ratio were used.

Were also added two classic financial ratios in bankruptcy prediction used by Altman; named in this work Altman x_1 and Altman x_2.

The following Table (I) described each one of the aforementioned ratios, according to NIFF financial statements retrieved for this work.

III. Modeling prediction

The models used to compare prediction performance on bankruptcy were: Decision tree, logistic regression and multilayer perceptron.

was constructed a grid search.

A. Train and test

We used 70% to train and 30% to test.

B. Hyperparameter tunning

C. Logistic regresion

D. Multilayer perceptron

E. Decision tree

-

F. Balance data

To handle the unbalanced data used undersampling

TABLE I: Financial ratios definition

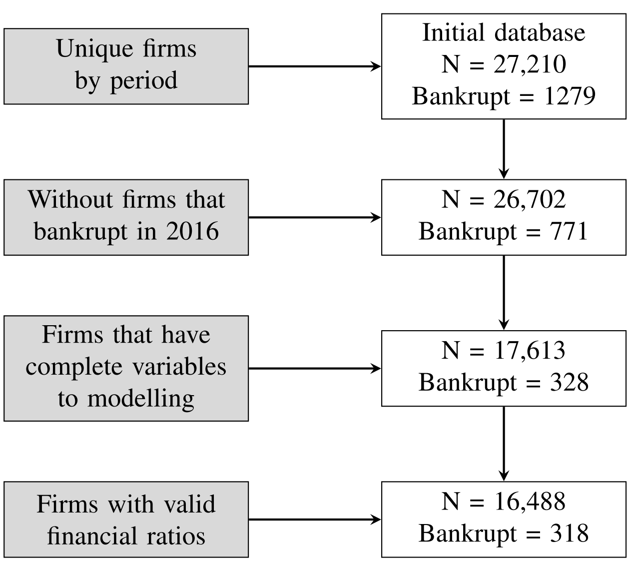
Ratio	Definition
Gross Profit Margin (GPM)	$\frac{\text{Gross profit}}{\text{Operating revenues}}$
Net Profit Margin (NPM)	$\frac{\text{Profit(Loss)}}{\text{Operating revenues}}$
Return On Equity (ROE)	$\frac{\text{Profit(Loss)}}{\text{Total equity}}$
Return On Assets (ROA)	$\frac{\text{Profit(Loss)}}{\text{Total assets}}$
Indebtedness Ratio (IR)	$\frac{\text{Total liabiliaties}}{\text{Total assets}}$
Debt Equity Ratio (DER)	$\frac{\text{Total liabiliaties}}{\text{Total equity}}$
Current Ratio (CR)	$\frac{\text{Total current assets}}{\text{Total current liabilities}}$
Ratio of Short-Term Liabilities (RSL)	$\frac{\text{Total current liabilities}}{\text{Total liabilities}}$
Altman x_1	$\frac{\text{Total current assets - Total current liabilities}}{\text{Total assets}}$
Altman x_2	$\frac{\text{Accumulated profits}}{\text{Total liabilities}}$

IV. Results

From a total sample of 27,210 firms, of which 1,279 were classified as bankrupt, only 16,488 of which 318 were bankrupt represented the final sample.

The following Figure (1) describe the process to obtain the final data to modelling.

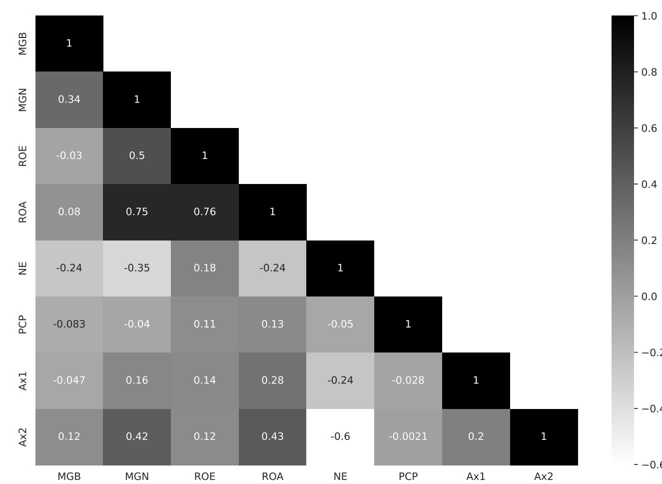
Fig. 1: Flowchart processing data



The Figure (2) show the spearman coefficient among numerical financial ratios.

The following Table II show the difference among the financial ratios of one year before of the bankrupt and the mean of the ratios by the period of activity in no-bankrupt firms. The results indicate that there are significative difference in all ratios, and no independence among the event and the sector.

Fig. 2: Spearman correlation in predictor variables



The following table (III) show us the performance of the models with the variables used to predict bankruptcy, the result indicate a lower performance to predict correctly the bankrupt firms.

The following Table (IV) shows that all models have major performance.

TABLE II: Year of bankrupt, financial ratios and sector among groups

				Grouped by event		
		Missing	Overall	No-bankrupt	Bankrupt	P-Value
n			16806	16488	318	
time-event, n (%)	0.0	0	16488 (98.11)	16488 (100.00)		¡0.001
	2017.0		99 (0.59)		99 (31.13)	
	2018.0		96 (0.57)		96 (30.19)	
	2019.0		123 (0.73)		123 (38.68)	
GPM, median [Q1,Q3]		0	0.31 [0.18,0.56]	0.32 [0.18,0.56]	0.26 [0.14,0.41]	¡0.001
NPM, median [Q1,Q3]		0	0.03 [0.00,0.08]	0.03 [0.00,0.08]	-0.02 [-0.18,0.03]	¡0.001
ROE, median [Q1,Q3]		0	0.07 [0.01,0.16]	0.07 [0.01,0.16]	0.01 [-0.18,0.09]	¡0.001
ROA, median [Q1,Q3]		0	0.03 [0.00,0.06]	0.03 [0.00,0.07]	-0.01 [-0.09,0.01]	¡0.001
IR, median [Q1,Q3]		0	0.53 [0.32,0.72]	0.52 [0.32,0.72]	0.74 [0.59,0.88]	¡0.001
RSL, median [Q1,Q3]		0	0.75 [0.45,0.98]	0.75 [0.45,0.98]	0.53 [0.28,0.87]	¡0.001
Ax1, median [Q1,Q3]		0	0.22 [0.04,0.43]	0.22 [0.04,0.43]	0.09 [-0.08,0.28]	¡0.001
Ax2, median [Q1,Q3]		0	0.19 [0.04,0.40]	0.19 [0.05,0.40]	0.04 [-0.12,0.18]	<0.001
Sector, n (%)	A	0	1012 (6.02)	990 (6.00)	22 (6.92)	¡0.001
	B		282 (1.68)	277 (1.68)	5 (1.57)	
	C		2735 (16.27)	2647 (16.05)	88 (27.67)	
	D		10 (0.06)	10 (0.06)		
	E		28 (0.17)	27 (0.16)	1 (0.31)	
	F		2238 (13.32)	2190 (13.28)	48 (15.09)	
	G		5288 (31.46)	5201 (31.54)	87 (27.36)	
	H		314 (1.87)	310 (1.88)	4 (1.26)	
	I		363 (2.16)	349 (2.12)	14 (4.40)	
	J		464 (2.76)	456 (2.77)	8 (2.52)	
	K		474 (2.82)	473 (2.87)	1 (0.31)	
	L		1449 (8.62)	1443 (8.75)	6 (1.89)	
	M		1142 (6.80)	1127 (6.84)	15 (4.72)	
	N		646 (3.84)	632 (3.83)	14 (4.40)	
	O		4 (0.02)	4 (0.02)		
	P		87 (0.52)	86 (0.52)	1 (0.31)	
	Q		64 (0.38)	63 (0.38)	1 (0.31)	
	R		88 (0.52)	86 (0.52)	2 (0.63)	
	S		115 (0.68)	114 (0.69)	1 (0.31)	
	U		3 (0.02)	3 (0.02)		

TABLE III: Models performance over imbalanced data

	Logistic Regression		Decision Tree		Multilayer Perceptron	
	No-Default	Default	No-Default	Default	No-default	Default
precision	0.98	0.15	0.98	0.06	0.98	0.00
recall	1.00	0.03	0.94	0.18	1.00	0.00
f1-score	0.99	0.05	0.96	0.09	0.99	0.00
support	4947.00	95.00	4947.00	95.00	4947.00	95.00

TABLE IV: Models performance over balanced data using undersampling

	Logistic Regression		Decision Tree		Multilayer Perceptron	
	No-Default	Default	No-Default	Default	No-default	Default
precision	0.68	0.72	0.71	0.64	0.72	0.76
recall	0.75	0.64	0.58	0.76	0.78	0.69
f1-score	0.71	0.68	0.64	0.70	0.75	0.73
support	96.00	95.00	96.00	95.00	96.00	95.00