Application of data mining techniques for understanding capital structure of Brazilian companies

R. A. M. Horta[1,2], H. M. Pires[3] & B. S. L. P. de Lima[2]
[1]*Finance and Accounting Department,*
Federal University of Juiz de Fora, Juiz de Fora, Brazil
[2]*Civil Engineering Program, Coppe,*
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[3]*Industrial Engineering Program, Coppe,*
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Abstract

The capital structure decision is related to the adoption of strategies for the choice of financing sources for the own capital and the thirty party capital. Numerous studies were developed aiming at clarifying the adoption of those strategies.

This paper aims to identify attributes that influence the decision of capital structure through an exploratory data analysis using records of Brazilian companies of open capital. The financial companies were not included in this study. Thus, it is expected to achieve new information on this process of knowledge discovery through statistical and data mining techniques, such as linear regression, robust regression and neural networks, in order to get a better assessment of the financing strategies employed by companies.

The data analysis was made in two stages employing as dependent variables the onerous debt and the equity degree of financial leverage. The database consists of companies on BOVESPA classified by economic sectors in the period from 1996 to 2006.

The results suggest that the explanatory variables of financial leverage are: economic sector; return on equity; return on assets; immediate liquidity; general liquidity and assets turnover.

Keywords: data mining, capital structure, attribute selection.

WIT Transactions on Information and Communication Technologies, Vol 42, © 2009 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA090111

1 Introduction

Capital structure choice is one of the main topics of study on corporate finance. The company can constitute this structure with countless combinations and the explaining factors of those combinations gave birth to the modern theory of capital structure.

The study of Modigliani and Miller [1] is considered the most important on indebtedness policy. This work represents the basis for the development of many studies that analyse the influence of introducing market imperfections into capital structure decisions. It is within this perspective that studies and researches are carried out based on the trade-off theory, agency cost theory, informational asymmetry theory as well as on the determinants of capital structure choice, thus emphasizing the influence of product/company nature and market competition.

The trade-off theory lies on the idea that indebtedness brings both benefits and costs to the company. The benefits arise essentially from the fiscal advantage obtained from interest deductibility. As for costs, they are related to company bankruptcy, more precisely, to financial difficulty costs. Therefore, the trade-off theory proposes the existence of an optimum capital structure that maximises company value by balancing indebtedness benefits and costs [2].

Afterwards, in addition to the financial difficulties-related costs, those related to agency costs [3], potential loss of fiscal advantages not resulting from indebtedness [4], tax difference between earnings from shares and from debt instruments are considered, and all those elements exert influence on the choices in terms of capital structure.

As for agency theory, it deals with the conflicts that take place in the organisations due to the different interests and attitudes among the different players involved. Those conflicts incur into costs that, as a consequence, influence on the capital structure choices, for the shareholders may be reluctant to pursue new capital, whether by means of new partners or by third capital.

Another important element in the explanation of company indebtedness decisions is the asymmetric information between managers and outside investors, to an extension that the former are better informed than the latter on the payback foreseen, risks involved, investment opportunities and company's operations, among other things.

By making use of a model that deals with the companies' issue-invest decision balance, Myers [5] show that information asymmetry associated with the fact that the managers serve in favour of the interests of the current shareholders generates situations in which managers reject good investment opportunities.

One of the most remarkable works in this line of research is by Titman and Wessels [6]. Taking advantage of some attributes, these authors pursue to identify their influence on the North American company indebtedness level. Among the attributes, it is worth highlighting here assets structure, company growth and industry size and classification.

For Harris and Raviv [7], the 'introduction in the economy of explicit models of private information has made possible various approaches to explain capital structure'. Those approaches are divided into two groups. In the first one, the choices related to company capital structure point to the external investors the information held back by the managers (the insiders). This group – represented by the idea of capital structure as information indicator – is divided into two subgroups: (i) level of indebtedness and (ii) models based on aversion to manager risk. Finally, the authors discuss the way through which the managers pursue third party or shareholders capital sends signals to the market about the company's conditions.

Miao [8] studies capital structure of high technology industry. For him, they are companies of which investments and operational decisions comprise assets that demand high investments and are affected by a fast technological devaluation. The financial structures reflect the relationship between investment paybacks, insolvency cost, technological risks, technological changes and tax policy.

Byoun [9] studied capital structures that conform themselves to the financial changing conditions due to a financial deficit/surplus.

The present article intends to investigate, by means of variables selected by data mining pre-processing techniques, the strategies of capital structure composition adopted by Brazilian companies for extracting description rules in the business scenario. It is employed three approaches to reduce the dimensionality of the data – filter, wrapper and principal components analysis.

2 Data analysis

The main objective of this section is to perform an attribute selection. This task is an important pre-processing phase in the knowledge discovery process. Attribute selection implies sample reduction without, however, causing loss of the characteristics of the sample, allowing the improvement of the results.

The first approach used is the filter approach that employs general types of the data to evaluate attributes and operate independently of any learning algorithm. The wrapper approach evaluates attributes by using accuracy estimates provided by the actual target learning algorithm.

2.1 Filter approach

The goal is to select a subset of attributes that preserves, as much as possible, the relevant information found in the entire set of attributes [10]. The idea is to filter irrelevant attributes according to some criterion before the learning process takes place.

The techniques employed as search methods for the best subset of attributes used in the filter approach are: Genetic Selection (GS) [11] and GreedyStepwise (SP).

In order to evaluate the subsets of attributes it is employed the Correlation-based Feature Selection (CFS). CFS is a heuristic that evaluates subsets of

attributes taking into account the usefulness of individual features for predicting the class along with the level of intercorrelation among them [12].

The correlation between attributes X and Y is computed by estimating their degree of association by using the symmetric uncertainty (SU) as follows:

$$SU(X,Y) = 2.0 * \left[\frac{H(X) + H(Y) - H(X,Y)}{H(X) + H(Y)} \right].$$

where H is the entropy function described in [13].The entropies are based on the probability associated with each attribute value. $H(X,Y)$ is the joint entropy of X and Y which is calculated from the joint probabilities of all combinations of X and Y values. SU lies between 0 and 1.

After computing the correlation matrix, CFS assigns high scores to subsets containing attributes that are highly correlated with the class and have low intercorrelation with each other. A merit of a subset S containing k features is defined as

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}},$$

where r_{cf} is the average feature-class correlation and r_{ff} is the average feature-feature intercorrelation.

2.2 Wrapper approach

This approach assesses the relevant attributes using accuracy estimates provided by pre-determined learning algorithms [10]. In this study, three pre-determined learning algorithms were used: Linear Regression (LR); Robust Regression (RR); and Multilayer Perceptron (MLP).

The learning algorithm that employs LR uses the Akaike Information Criterion (AIC) to evaluate the quality of the fit in addition to eliminate the collinearity between variables. RR was applied in order to overcome some limitations of LR mainly with the presence of outliers. A feedforward artificial neural network, MLP, is also employed as a learning algorithm in the wrapper approach.

2.3 Principal Component Analysis (PCA)

Principal Component Analysis is a statistical technique that can reduce the dimensionality of data as a by-product of transforming the original attribute space. The basic idea of the method is to transform the correlated p variables into $k<p$ non-correlated linear combinations.

3 Methodology

This empirical research was descriptive and quantitative in nature, comprising the companies rated by sector at BOVESPA (São Paulo Stock Exchange).

3.1 Market-value versus book-value data

Agarwal and Taffler [14] compared results of tests using market-value and book-value and showed that in fact, accounting data-based models significantly produce greater economical benefits than those based on the market. Blöchlinger and Leippold [15] observed that the differences related to errors are economically significant in favour of accounting data. Thus, in this work, it was adopted the book value.

3.2 The dataset

The dataset consists of 175 Brazilian open-capital companies during the period from 1996 to 2006 provided by BOVESPA.

Of the 175 companies studied 28 were classified at CVM (Securities and Exchange Commission of Brazil) as insolvent (bankrupted, under creditor's arrangement or judicial recovery), and the remaining 147 companies were classified as solvent companies. The preparation of a solvent company group followed the same sector-based classification of the insolvent companies, in addition to the similarity of the total assets size.

In the sample composition, economical-financial data were obtained on the solvent companies from the last ten years and on insolvent companies from the last five years. The first year of the insolvent companies is related to the year in which the company broke. Hence, the data bank adds up 1.610 instances, with 1.470 related to solvent companies and 140 to insolvent companies.

The database consists of twenty-two economical-financial indicators presented on Appendix 1. Consolidated balance sheet data and financial institutions were not included, being the goal to study the companies separately. The economical sector of each company was also listed according to BOVESPA rating in 2007. Appendix 2 shows details of sectors.

It was employed the software WEKA version 3.4.8 [13]. A ten-fold cross-validation is used for accuracy estimation.

3.3 Dependent variables

Two variables were chosen in order to represent the indebtedness policies adopted by the companies for the financial balance of the period studied

The first variable depending on capital structure is the Onerous Indebtedness on Net Equity – EOPL. More emphasis is given to the long-term funds in this EOPL variable, since it is assumed that the current liabilities resources first aim at meeting the seasonal financial needs of the companies and not at financing the demand for permanent resources. Permanent resources are expenses in which the entity has to necessarily incur to in order to keep itself in continuity conditions, namely expenses with suppliers, employees' payroll, taxes and fiscal, social security and labour provisions. This variable evidences the relationship level existing between the onerous financial requirements to leverage the assets in relation to the entity own capital. EOPL was chosen as a dependent variable in

order to obtain through a data analysis an explanation of the companies choice of capital structure.

The second dependent variable is defined as financial leverage level – GAF. This variable is the ratio between the payback rate on Net Equity and on Assets. GAF can measure the effect of the indebtedness level of the company financial structure in addition to measure whether or not the company capital structure is benefiting its shareholders. Financial leverage measures the capacity that the company has to manage own and/or third parties' resources and, consequently, maximise shareholders' profits.

Table 1: Summary table of the variables correlation.

	EOPL						GAF					
	Wrapper			*Filter*		PCA	*Wrapper*			*Filter*		PCA
	RR	LR	MLP	GS	SP		RR	LR	MLP	GS	SP	
ES	X	X	X	X	X	X	X	X	X	X	X	X
ROE	X		X	X	X	X	X		X	X	X	X
ROA	X	X	X	X	X	X		X	X	X	X	X
RTA	X					X	X					X
ROI		X				X	X			X	X	X
GA		X		X	X		X	X		X	X	
GM						X						X
OM			X			X			X			X
NM			X			X			X			X
EBIT			X			X			X			X
EBTIDA			X			X			X			X
IMCP				X	X		X		X	X	X	
IL				X	X	X		X		X	X	X
DL						X		X				X
CL						X	X	X				X
GL		X		X	X	X		X	X	X	X	X
TERFIN										X	X	

4 Case study

In order to fully understand how some variables have impact on the dependent variables it was performed the data analysis employing the techniques listed in section 2. It is important to obtain the structure of the description that is learned by the algorithms in terms of the most important variables and how they relate to the numeric predictions which are the dependent variables in this case study.

Table 1 shows 17 independent variables and their relation to the two dependent variables, EOPL and GAF, applying wrapper approach, filter approach and PCA. It would be interesting to acquire the smallest number of variables or factors to describe this relationship.

The results of PCA in Table 1 showed that a great number of principal components account for 86 % of the variance in the dataset. This technique obtained the smallest reduction in the data.

Table 2 show the quality measures values (R^2) of the previous numeric prediction with the purpose of evaluating the performance of those analyses. It was used the basic principles for performance evaluation using an independent

test set rather than the training set with holdout and 10-fold cross validation methods.

Table 2: Performance measure of wrapper approach.

	GAF		EOPL	
Learning algorithm	R^2	R^2 w/cv	R^2	R^2 w/cv
LR	0,2842	0,2288	0,4916	0,3782
RR	0,2575	0,0944	0,0241	0,1238
MLP	0,3893	0,1981	0,7736	0,3765

5 Result analysis

The chosen learning algorithms employed in the wrapper approach were LR, RR and MLP since they work more naturally with numeric prediction problems

Observing Table 1 it is possible to have a good insight of the most important variables selected by the attribute selection techniques using *majority voting*. Some variables have significant impact on the dependent variables such as Economic Sector (ES). This variable is present in all attribute selection techniques. The results suggest that companies of the same economic sector present capital structures with considerable resemblances.

Two profitability variables, as return on equity (ROE) and return on assets (ROA) were well chosen by the selection techniques, which suggest that profitability is preponderant for the company financial strategy composition. Other variables as assets turnover (GA), immediate liquidity (IL), and general liquidity (GL) were also selected.

For Ross *et al*. [16], immediate liquidity is defined as a short-term solvency. Schrickel [17] defines it as the company's capacity to settle debts, being suitable for a situation of a continuous business, handling the convertibility rate of current assets and current liabilities. This convertibility rate, a mirror of the company's operating cycle, characterises the liquidity condition. LG variable evidences the company's capacity to pay for all its short commitments. In fact, this variable does not express a liquidity condition, but solvency. For the company, solvency translates into the capacity to liquidate its commitments on time. The presence of this variable demonstrates the indebtedness strategy oriented also to solvency capacity, i.e., it is related to the existence of a sufficient amount of assets capable of offering, at all times, suitable protection and renewing the commitments made with existing third parties. The results suggest that the Brazilian companies give importance to their solvency level allied to liquidity, trying to prevent them, in adverse economical situations, from largely compromising the operating activities and its continuity.

6 Conclusions and future studies

The main objective of this study was a preliminary understanding of the influence of some finance variables on the decision of capital structure of

Brazilian companies of open capital. This task was effectively performed by a data analysis employing some statistical and data mining techniques.

The most important attributes were selected using a dataset composed of indicators obtained from the companies' financial statements. This dataset was used exclusively in order to provide conclusions from real and legal sources and it was formulated to observe accounting standards, which go through further audit techniques to test and validate the data stated therein.

The variables present in the results that are considered significant were immediate liquidity, general liquidity, economic sector, assets turnover, return on equity and return on assets.

In future studies of this work, it is important to develop more detailed analyses on the variables selected, business environment, the companies and influences exerted by the economic sectors in the strategy adopted for the indebtedness policy.

Appendix 1 – Economical-financial rates

Current Liquidity (CL), Dry Liquidity (DL), Immediate Liquidity (IL), general liquidity (GL), onerous indebtedness on net equity (EOPL), financial leverage level (GAF), fixed assets of permanent resources (IMCP), gross margin (GM), operating margin (OM), net margin (NM), assets turnover (GA), return on assets (ROA), return on equity (ROE), return on operational assets (ROI), financial thermometer –TERFIN, Dupont adapted model (RTA), earnings before interest and taxes (EBIT), earnings before interest and taxes, depreciation/depletion and amortization (EBTIDA).

Appendix 2 – Economic Sectors (ES) –BOVESPA 2007

Industrial assets, Construction and transportation, Cyclic consumption, Non-cyclic consumption, Basic materials, Information technology, Telecommunications.

References

[1] MODIGLIANI, Franco; MILLER, Merton H. The cost of capital, corporate finance and the theory of investment. American Economic Review, v. 48, n. 3, p. 261-97, June 1958.

[2] BRADLEY, Michael; JARREL, Gregg A; KIM, E. Han. On the existence of an optimal capital structure: theory and evidence. The Journal of Finance, v. 39, n. 2, p. 857-878, July1984.

[3] JENSEN, Michael C.; MECKLING, William H. Theory of the firm: managerial behavior, agency costs, and ownership structure. Journal of Financial Economics**,** v. 3, n. 4, Oct. 1976.

[4] DeANGELO, Harry; MASULIS, Ronald W. Leverage and dividend irrelevancy under corporate and personal taxation. The Journal of Finance, v. 36, June 1980.

[5] MYERS, Stewart C.. The capital structure puzzle. The Journal of Finance, v. 39, n. 3, July 1984.
[6] TITMAN, Sheridan and WESSELS, Roberto. The determinants of capital structure choice. Journal of Finance, v. 43, p. 1-19, 1988.
[7] HARRIS, Milton; RAVIV, Artur. The theory of capital structure. The Journal of Finance, v. 46, n. 1.
[8] MIAO, J. Optimal Capital Structure and Industry Dynamics. The Journal of Finance . Vol. LX. No. 6. December 2005.
[9] BYOUN, Soku. How and When Do Firms Adjust Their Capital Structures toward Targets? The Journal of Finance. Vol. LXIII. No. 6. December 2008.
[10] FREITAS A. A. Data mining and knowledge discovery with evolutionary algorithms. Springer-Verlag Berlin Heidelberg, New York, 1998.
[11] GOLDBERG, David. Genetic Algorithms in Search, Optimization, and Machine Learning (Hardcover), 1989.
[12] HALL M. A. and HOLME G. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, 2003.
[13] WITTEN, I.H., FRANK E. Data Mining: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems, 2nd ed. 2005.
[14] AGARWAL, Vineet; TAFFLER, Richard. Comparing the performance of market-based and accounting-based bankruptcy prediction models. Journal of Banking & Finance 32 (2008).
[15] BLÖCHLINGER, A., LEIPPOLD, M., 2006. Economic benefit of powerful credit scoring. Journal of Banking and Finance 30, 851-873.
[16] ROSS, Stephen A.; JORDAN, Bradford D.; WESTERFIELD, Randolph W. Administração financeira. 8ª São Paulo: Ed. McGraw-Hill Interamericana, 2008.
[17] SCHRICKEL, Wolfgang K. Demonstrações Financeiras. 2ª Ed. São Paulo: Atlas, 1999.