Discriminant analysis in a credit scoring model

G. MIRCEA, M. PIRTEA, M. NEAMȚU, S. BĂZĂVAN
Faculty of Economics and Business Administration
West University of Timisoara, Romania
ROMANIA
gabriela.mircea@feaa.uvt.ro, marilen.pirtea@feaa.uvt.ro,
mihaela.neamtu@feaa.uvt.ro, sandra.bazavan@gmail.com

Abstract: - The purpose of this paper is to define a specific credit score model, based on discriminant analysis in order to complete financial diagnoses on particular predefined classes. The model is built based on a set of observations for which the classes are known. The classes in this paper are made of companies with certain characteristics which reflect the creditworthiness of that entity.

Key-Words: - **c**redit scoring, discriminant analysis, dicriminant indicators, risk assesment

1 Introduction

The available literature about credit scoring is mainly studying the evolution of financial indicators for a certain number of companies, which have failed or continued their activity during the analyzed period. The failure as well as the success of the management structure is being assessed by a particular indicator known as cutting score, which is defined as a linear combination of a few main financial indicators.

The scoring models represent a way of identify, quantify and control the corporate risk of bankruptcy [1]. Their multidimensional character follows a financial diagnosis of the entity and allows a relevant ranking of the companies, considering some financial indicators which are integrated in a score function.

The obtained results cannot be extended from one class of companies to another, due to the fact that the construction of the cutting score indicator is based upon specific branches and do not have a wider consideration. The purpose of our analyses is to find a synthetic indicator that suits to a random number of companies, therefore to achieve a value of its own as a reference for the status of the analyzed company.

At the moment there is no universal scoring model that could be used by all the financial institutions, due to the fact that each institution preserves its strategy in dealing with the customers [9]. The scoring model in this paper is based on discriminant analysis and it is pointed in the usage of the bank, by creating a tool that corresponds to random companies analyzed simultaneously.

We assume we have a group of companies called G which is formed of two distinct subgroups G1 and G2, each representing one of the two possible states: running order and bankruptcy. These two possible states are defined by a number of g independent financial indicators which simultaneously influence the progress of the companies, in terms of decreasing or growth.

2 Credit Scoring Fundamentals

We assume that a random bank has access to information about its customers, regarding both the good payers (reimbursing loan without problems) and the bad payers (who had problems with repayment over time). This information may relate to age, salary, social status, job stability and other reimbursement problems of individuals and to financial statements of legal persons. In this paper there will be taken in consideration only two indicators, but the algorithm itself is expandable to as many variables a bank cares for analyze.

When a new customer is applying for a loan, the bank must decide whether to grant him or not the requested loan by applying a discrimination rule. As a result of this process, the applicant will receive a score which classifies the application in one of the existing categories (*e.g.* bad payers, good payers). The discrimination rule offers support for decision of granting or not granting a loan, by attending at the background of the applicant and providing the required risk assessment.

We assume we want to define a credit score model for a group G of 14 companies with different characteristics and distinct values for two basic

ISBN: 978-1-61804-028-2

financial metrics, liquidity and debt to equity.

Therefore, we take in consideration the company's ability to pay off its short-terms debts obligations as well as the proportion of equity and debt the company is using to finance its assets.

We define the financial indicators that are taken in consideration for applying the algorithm, as it follows:

$$Dept\ of\ \ equity = \frac{Total\ liabilities}{Shareholders\ \ Equity} \quad (1)$$

$$Liquidity = \frac{Liquid\ \ assets}{Short\ Term\ \ Debts} \quad (2)$$

If this were to increase earnings by a greater amount than the debt cost (interest), then the shareholders benefit as more earnings are being spread among the same amount of shareholders. However, the cost of this debt financing may outweigh the return that the company generates on the debt through investment and business activities and become too much for the company to handle[1].

A high debt/equity ratio generally means that a company has been aggressive in financing its growth with debt, which can result in volatile earnings as a result of the additional interest expense. A higher value of the liquidity ratio denotes a larger margin of safety that the company possesses to cover short-term debts [2].

2.1 Discriminant Analysis Algorithm

When there is a case that requires a solution based on problem of discrimination, then it is automatically indicated a categorical type of variable as a reply. These variables place individuals literally into categories, and cannot be quantified in a meaningful way.

It is assumed that the data (for the variables) represent a sample from a multivariate normal distribution [5]. There can be examined whether or not variables are normally distributed with histograms of frequency distributions. However, note that violations of the normality assumption are usually not fatal, meaning, that the resultant significance tests etc. are still trustworthy [7]. There can be used specific tests for normality in addition to graphs.

In stepwise discriminant function analysis, a model of discrimination is built step-by-step. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the discrimination between groups [7]. That variable will then be included in the model, and the process starts again.

One can also step backwards; in that case all variables are included in the model and then, at each step, the variable that contributes least to the prediction of group membership is eliminated. Thus, as the result of a successful discriminant function analysis, one would only keep the "important" variables in the model, that is, those variables that contribute the most to the discrimination between groups [7].

We assume that the categorical variable defines a number of q available categories, so the sample of n entities (14 companies in our study) will be grouped in a number of q categories based on specific characteristics.

For a better representation, we note the vector of the indicators above with $x = (x_1, x_2, ..., x_g)$. Furthermore, by applying the same mechanism we can establish a vector of indicators for each of the subgroups G1 and G2 as it follows: $x_1 = (x_{11}, x_{21}, ..., x_{g1})$, $x_2 = (x_{12}, x_{22}, ..., x_{g2})$.

The first step in analyzing multivariate data is computing the mean vector and the variance-covariance matrix for each subgroup and than for the entire group. The mean and the variance-covariance matrix are denoted by symbols μ respectively β as following:

(μ_1, β_1) defines vector x_1,

(μ_2, β_2) defines vector x_2 and

$(\mu_1, \mu_2, \beta_1, \beta_2)$ defines vector x .

Step1. The matrix $X(n \times p)$ is defined by a number of n entities and g measured variables, both mentioned above. This matrix can be interpreted either line by line which case the interpretation leads to relevant data about the n entities, or column by column regarding information about the g measured variables.

Each entity of the n entities of the analyzed sample corresponds to a line in the matrix, meaning a vector containing a number of g elements (variables or indicators) will be written as it follows: $x_i = (x_{i1}, x_{i2}, ..., x_{ig})$.

Each variable of the g variables of the analyzed sample corresponds to a column in the matrix, meaning a vector containing a number of n entities will be written as it follows:

$x_j^n = (x_{1j}, x_{2j}, ..., x_{nj})$.

Step2. We define the vector *m* by its coordinates known as centroid, containing each mean of the *g* variables as it follows: $m=(m_1,m_2,...,m_g)$.

Step3. We define vector s as the vector of the standard deviations as it follows: $s=(s_1,s_2,...,s_g)$.

Step4. We define the estimated variance-covariance matrix for the *g* number of variables.

Step5. We define the vector containing each mean of the variables connected with the *q* number of categories. This particular vector is called the centroid of the category and for a random category called *c*, it can be written as it follows: $m^c=(m_1^c,m_2^c,...,m_g^c)$.

Step6. We define the covariance matrix of the *g* variables. By projecting the coordinates $(x_{k1},x_{k2},...,x_{kp})$ of a random *k* entity on the axe Δ with coordinate $u=(u_1,u_2,...,u_g)$ it is obtained the following value:

$$c_k=x_{k1}\times u_1+x_{k2}\times u_2+...+x_{kg}\times u_g \quad (3)$$

The value of c_k is called the score of the axe and represents the distance between the projection of entity *k* and the centroid of the vector *m* [7].

The synthetic indicator we want to obtain is usually expressed by a linear combination of the *g* indicators as the score mentioned above. The synthetic indicator will be compared to the value of z_c called cutting score which has the same architecture of the score itself, but it has to be predefined for the analyzed sample of entities.

The purpose of a discriminant technique is to find the axe Δ for which the discrimination of the projection is maxim.

The classification of the *n* entities upon the discriminant function is possible by referring to equation (1). Hence, we will obtain the values c^1,c^2 as projections of the centroids of the two categories on the axe. Therefore, the cutting score of the entities on the axe is defined as it follows:

$$c_{cs}=\frac{\eta_1\times c^1+\eta_2\times c^2}{\eta_1+\eta_2} \quad (4)$$

A random entity which achieves a score c_k has two possibilities regarding its position to the cutting score c_{cs}, respectively above it or beyond it. Each option classifies the entity in one of the two mentioned categories.

The success rate of discrimination is defined by the formula:

$$r_s=\frac{\eta_{11}+\eta_{22}}{\eta_1+\eta_2} \quad (5)$$

We have noted the spread of the entities in the class (group/subgroup) with the symbol η.

Initial group	Number of entities in the initial group	Group after classification	
		1	2
1	η_1	η_{11}	η_{12}
2	η_2	η_{21}	η_{22}

Table 2: Success rate of discrimination

A similar process of classification for two categories with the same distribution would lead to a success rate of discrimination of 50%. Therefore, the value difference between r_s and 50% is considered an indicator for the quality of the discrimination.

It is assumed that the variance/covariance matrices of variables are homogeneous across groups. Minor deviations are not that important; however, before accepting final conclusions for an important study it is probably a good idea to review the within-groups variances and correlation matrices. When in doubt, it is recommended to try re-running the analyses excluding one or two groups that are of less interest [7].

3 Applied scoring architecture

In this paper we will present a computer application that follows the algorithm mentioned above, using the Visual Basic programming environment.

The proper functioning of a financial institution is represented by the informational system based on a viable architecture that ensures quick and secured access to information.

The versatility of the computing platforms provides agility to control risk yet accelerate projects development [10]. A greater range of strategic information initiatives takes place by filling functionality around data quality, data services and business based on Service-oriented architecture. Service-oriented architecture (SOA) is a flexible set of design principles used during the

phases of systems development and integration in computing. A system based on a SOA will package functionality as a suite of interoperable services that can be used within multiple, separate systems from several business domains [7].

The challenge for all banks is not only to create a centre of excellence with established international standards of communication, but also to reconstruct and automate their business processes to maximize efficiency.

The IT systems that currently support each service require review and extension. The technology used must be future-proofed to suit integration of existing and future development platforms, without meaning core system replacement. Within a bank's legacy systems reside vital components of the organization's competitive edge, the mission-critical processes and systems that form the heart of the enterprise [2]. They may require rationalization, documentation and better understanding, for the benefit of extracting more value – but nevertheless they exist, and have been bought and paid for, and have proven reliability; processing billions of transactions per day across the globe. Core services should be individually identifiable and re-usable, such that systems development is far easier and quicker. In this way construction and maintenance costs are significantly reduced [4]. Service Oriented Architecture SOA is a key technology concept to achieve this level of re-use and avoids the extreme cost and risk of complete systems replacement.

SOA is one of the first steps in addressing the necessity to modernize business capability using technology as opposed to making the capability fit within the constraints of technology [7].

In Figure 1 we present the main form of the application that collects data and transforms it into relevant information for further decisions.

The user introduces the name of the company or any kind of information that particularly defines the analyzed entity, as well as the necessary financial fields from the customer's application. These fields are represented by the total of liabilities, the shareholders equity, the liquid assets and the short term debts. When introduced, the application automatically determines the two indicators that we take in consideration, respectively debt to equity and the liquidity of the company.

The user has the possibility of analyzing one or more companies, each time by clicking the Append to Sample button. This will send all the data that has been introduced so far into a vector called Initial Sample Data.

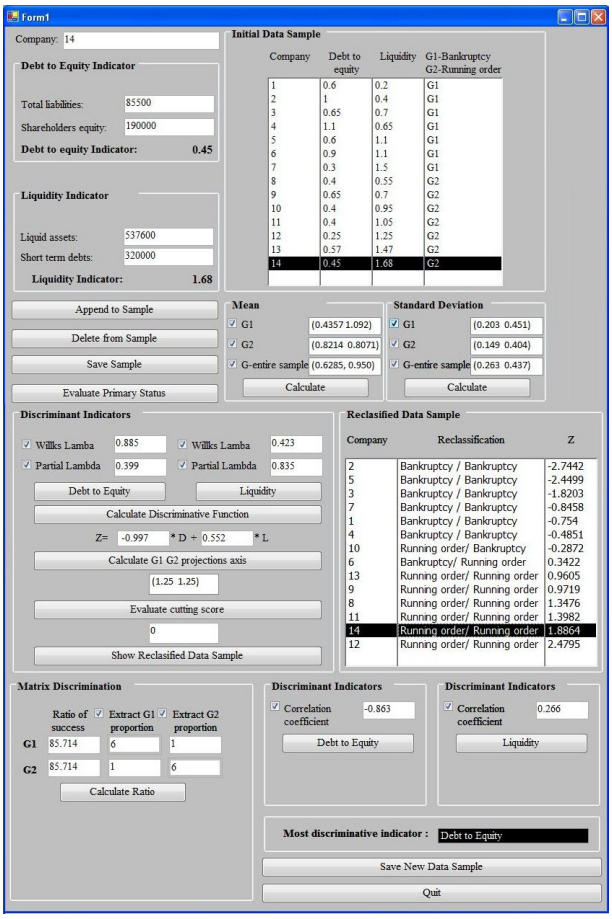

Figure 1: Main Form of the application

The user has the possibility to save, as well as to erase data from the sample. However, at this moment the Initial Sample Data vector is not fully completed, unless the user clicks on Evaluate Primary Status button first. By clicking this button, the application does a primary classification in one of the subgroups, G1-Bankruptcy or C2- Running order, following a pattern that is usually used in basic credit scoring models.

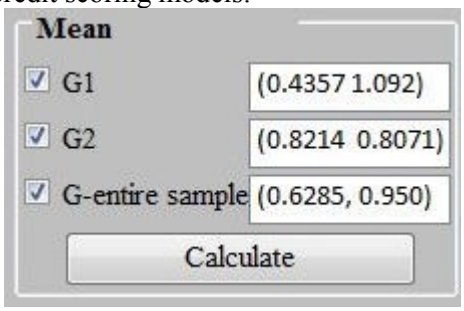

Figure 2: Calculating the mean

Step1. In order to continue the algorithm, the application determines the specific mean for each subgroup as well as for the entire group, by clicking the Calculate button in the group called Mean.

When selecting one or more options in the group Mean, the application transforms data in the

Initial Data Sample vector and displays it into each corresponding textbox.

The algorithm has calculated the mean for each subgroup, respectively *mG2* = (0.4357, 1.092) and m*G1* = (0.8214, 0.8071), as well as for the entire group: *m* = (0.6285, 0.950)

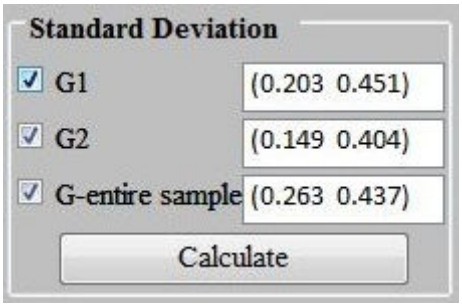

Figure 3: Calculating the standard deviation

Step2. In order to continue the algorithm, the application determines the specific mean for each subgroup as well as for the entire group, by clicking the Calculate button in the group called Standard Deviation.

When selecting one or more options in the group Standard Deviation, the application transforms data in the Initial Data Sample vector and displays it into each corresponding textbox.

The algorithm determines the standard deviation for each subgroup, respectively *sG2* = (0.149, 0.404) and *sG1* = (0.203, 0.451), as well as for the entire group *s* = (0.263, 0.437).

Step3. The algorithm determines the variance-covariance matrix for the subgroups $W_{G1G2} = \begin{pmatrix} 0.0273 & 0.0168 \\ 0.0168 & 0.1575 \end{pmatrix}$ and for the entire group $W = \begin{pmatrix} 0.0645 & -0.0107 \\ -0.0107 & 0.1779 \end{pmatrix}$.

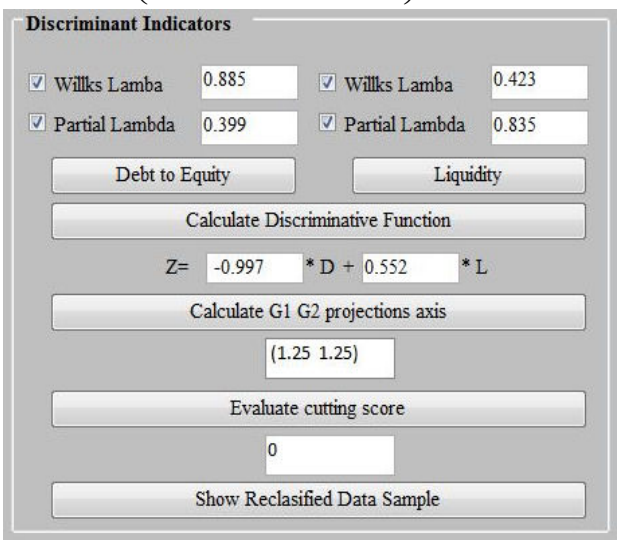

Figure 4: Discriminant indicators

The algorithm determines as well as the correlation matrix for the subgroups $V_{G1G2} = \begin{pmatrix} 1 & 0.256 \\ 0.256 & 1 \end{pmatrix}$ and for the entire group $V = \begin{pmatrix} 1 & -0.100 \\ -0.100 & 1 \end{pmatrix}$.

Step4. In order to continue the algorithm, the application has to calculate the Willks Lambda and the Partial Lambda for each of the two indicators taken into consideration.

The value of the main diagonal elements proportionally leads on the success of the discriminant ratio.

Step5. The application will determine the standard discriminant after the Lambda functions have been established, as following:

$$z = -0.997 * D + 0.552 * L \qquad (6)$$

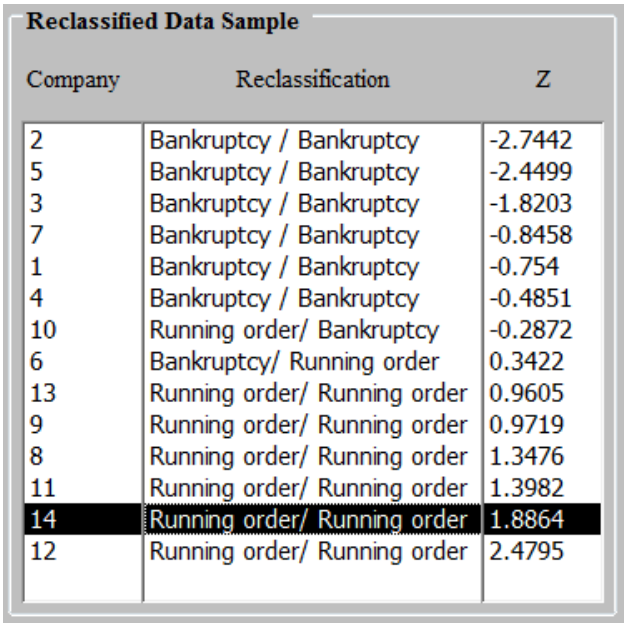

Figure 4: Reclassified Data Sample

The centroids of the subgroups G1- Banktrupcy and G2- Running order are projected on the axe Δ, using the coordinates (1.25, 1.25), which the application reveals after the user click on the Calculate G1, G2 projection axis button. Therefore, the application establishes the value of $z_c = 0$ the cutting score in current analysis.

The user has now the possibility of reclassifying the Initial Data Sample, which is basically one of the most important aspects of our scoring model, because it minimizes risk by offering a more accurate view upon the current analyze.

ISBN: 978-1-61804-028-2

Step6. We want to determine the class of an observation based on a set of variables known as predictors or input variables. These discriminant functions are used to predict the class of a new observation with unknown class.

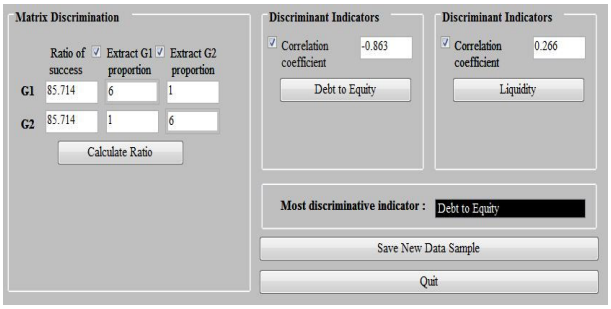

Figure 5: Discriminant Indicators

Figure 5 presents the ascending scorings of the companies that resulted after the applying the discriminant function (6). The algorithm analyzes the value of the most discriminative indicator which turns to be debt to equity indicator, due to its highest value for the F-statistic. The F-statistic represents the ratio between the spread of the class (subgroups G1- Bankruptcy and G2- Running order) and the spread inside the class.

4 Conclusion

The discriminant analysis represents an effective method for multivariate data analysis, often being use to extract relevant information from large and heterogeneous amounts of data. As a technique for classifying a set of observations into predefined classes, the discriminant analysis highlights their similarities and differences between them, creating an important advantage in describing the variability of a data set. Therefore, the method reduces the number of dimensions, without a significant loss of information. Probably the most common application of discriminant function analysis is to include many measures in the study, in order to determine the ones that discriminate between groups [7].

The loan processing has rapidly increased in speed due to scoring systems. Rather than perform lengthy credit investigation, creditors and other lenders are able to access credit scores to determine credit risks [8]. Due to the nature of its business, risk management is inherent to financial industry. In banking there is an ever present risk of payment default, fraud, theft, identity theft, and operational risk connected with internal procedures and processes [6]. Traditionally, the banking system focuses on IT development in Back Office regarding operational aspects and production. The next stage attends the compliance with local and international laws and regulations in the industry such as anti-money-laundry. After this stage there are taken in consideration aspects of customer management systems and distribution, meaning the Front Office. These represent major opportunities of growth, because the first attracts customers and the second helps their effective management, by providing differentiation in the market.

Service orientation represents a construction method rather than a technology, being applied particularly to one system when it needs to offer functions to other systems [3].

The strategy of cost reduction by avoiding unnecessary risks represents an opportunity of the financial sector to compete and maintain profitability. It makes business sense for an industry that recognizes the importance of exploiting competitive advantage through technology to consider an approach that allows them to identify, upgrade and re-use existing legacy assets, while also taking advantage of new technologies.

Applying service orientation is a way of improving the business flexibility and agility required by the competitive environment, providing an ensemble that is meaningful to the business and hides the technical components.

References:

[1] B. Baesens, T. Van Gestel, S. Viaene, M., *Benchmarking state-of-the-art classification algorithms for credit scoring,* Journal of the Operational Research Society, 2003;
[2] G. Fernandez, *Data Mining Using SAS Applications*, Charman & Hall, 2003;
[3] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition,*2005;
[4] L. C. Thomas, D. B. Edelman, and J.N. Crook, *Credit Scoring and its Applications,* SIAM, Philadelphia, USA, 2002;
[5] D. Power, *Proceedings of Informing Science and IT Education,* Santa Rosa: The Informing Science Institute. Retrieved December 1, 2005;
[6] G. Radonić, *A Review Of Business Intelligence Approaches To Key Business Factors In Banking*, Journal of Knowledge Management Practice, Vol. 8, SI 1, May 2007;
[7] http://revistaie.ase.ro/content/9/odagescu.pdf
[8] http://www.creditscoring.com/creditscore/distribution.html
[9] http://www.statsoft.com/textbook/discriminant-function-analysis/
[10] http://en.wikipedia.org/wiki/Banking_software