

THE MOST RELEVANT VARIABLES TO SUPPORT RISK ANALYSTS FOR LOAN DECISIONS: AN EMPIRICAL STUDY

CHARLO, María José*

Abstract

This paper presents an empirical study about risk analysis using several technologies. We have investigated two fundamental questions: we have analyzed the results obtained using each specific technology (discriminant analysis, logistic regression, and an artificial intelligence technology) and we have pointed out the most relevant variables for this kind of decision making. The percentage of error when using an artificial intelligence technology allows us to conclude that these intelligent systems are a good support in decision-making for risk analysts from banking entities. The most important variable using an artificial intelligence technology is the firm's economic situation, followed by firm rating and by firm image.

Keywords: Credit risk analysis, artificial intelligence, CBR, discriminant analysis, logistic regression.

JEL Classification: C40, D81, G20, G24,

1. Introduction

A great variety of mathematical models have been used in the analysis of credit exposure, mainly discriminant analysis and logistic regression. Other techniques belonging to artificial intelligence such as expert systems, neural nets and case-based reasoning systems have also been used. The purpose of all of them is to support risk analysts in their proposal for possible loan decisions.

Banking entity analysts, usually heading a group, report on whether a loan should be granted or not, and if they are entitled to it, make a decision. In his work analysts rely on information both on the firm and on the loan itself. This information will be processed in different ways depending upon each banking entity.

This paper first outlines the database chosen for the empirical study, explaining briefly the workings of the different models used. Then, the different technologies described above are applied in order to know the findings of each one.

2. Data collection

A sufficiently large sample is necessary to perform the different tests in order to be able to draw some conclusions. Processed information is both accounting and non-accounting, quantitative and qualitative; this information is the same that was originally analysed by the banking entity's risk analyst; and therefore, being real information, it suits perfectly the research study's objectives. It is information on firms that applied for loans to banking entities. It includes granted and rejected applications; likewise

* María José Charlo, Faculty of Economics and Business, University of Sevilla (Spain)

applications granted under certain conditions were also included. The database has been borrowed from a questionnaire developed by Nuñez Garcia (1997) which includes sixteen explanatory variables and one result variable: the analyst's proposal as to grant, reject or grant under certain conditions a given loan. The sixteen variables selected to obtain the sample, which refer to the firm and the loan it is applying for, are as follows:

1. Does the firm present a net worth statement?
2. Does the firm have a double set of books?
3. Does it meet legal requirements in terms of taxes, social security and filing of accounts at the Mercantile Register?
4. Does it present any title or corporate deed?
5. Is there any negative information in registers such as the Dad Debt Register, etc.?
6. Opinion drawn from personal interview
7. Financial situation – current – is it good, average or bad?
8. Financial situation – based on evaluation – improves, stays or deteriorates?
9. Economic situation - current – is it good, average or bad?
10. Economic situation – based on evaluation – improves, stays or deteriorates?
11. Management appraisal
12. Industry classification
13. Firm rating
14. Risk rating
15. Are the funds allocated to an appropriate purpose?
16. Does the firm have a positive image?

Initially this database included 169 loans and it was extended by means of interviews to risk analysts to a total of 279 loans corresponding to different firms. The resulting sample is appropriate for the intended research study because it shows the behaviour of analysts when facing a certain risk situations. It is also necessary to note that it includes loans falling within the three categories mentioned above. All this allows for comparing between the proposal made by the banking entity's risk analyst and the results obtained by the different methods chosen to conduct such analysis.

Statistical techniques will be applied using the SPSS version 10.0 statistical package, which will allow for easy and automatic import of the sample in EXCEL format. The case-based reasoning technology tests will be conducted by means of the ReMindTM Package.

3. Discriminant analysis, logistic regression and case-based reasoning systems

When carrying out the empirical work by means of some statistical techniques the first problem faced is to choose the most convenient ones among them. In trying to select the most appropriate technique for the intended analysis it must be taken into account that there is no metrical risk measure. However, we do know that a given firm is in a good or a bad situation as to be able to make a decision on whether to grant or not the loan applied for.

In this situation in which the dependent variable of the study (the analyst's proposal on the appropriateness of granting a loan to a firm or not) is a categorical one, discriminant analysis and logistic regression are appropriate statistical techniques. Furthermore, both of them are capable of dealing both with two groups and with multiple groups (three or more); although logistic regression, also known as logit analysis is restricted in its basic form to two groups, in alternative formulations it can consider more than two groups.

As pointed out by Mar-Molinero and Serrano-Cinca (2001), discriminant analysis and logit analysis are the most widely used mathematical models by researchers focusing on firm bankruptcies, an aspect which is closely related to risk analysis, as indicated by Mora Enguidanos (1995, p. 290) or Gabas Trigo (1997, p.16): "banking entities by just obtaining the same information may assess financial solvency levels and bankruptcy probabilities".

Thus, both statistical techniques are suited for this analysis, in which the sample includes a dependent variable with three possible values: to grant the loan, to reject it, or to grant it under certain conditions.

Discriminant analysis is a technique that lets us know which variables are relevant to classify data, by obtaining a Z indicator, which discriminates among groups. Discrimination is performed by establishing the weights of the theoretical value for each variable, i.e. a linear combination of two (or more) independent variables that discriminates better among the groups that have been defined. This analysis starts from a series of restrictive hypotheses: the assumption of a multivariate normal distribution of variables and equal matrices of variances and covariances.

Logistic regression is an alternative to discriminant analysis since it is less influenced by variance – covariance differences, and furthermore, it can manage categorical independent variables more easily. Logistic regression is a special type of regression used to predict and explain a categorical variable. The theoretical value represents a multivariate relationship that indicates the relative influence of the predictor variable.

In order to compare the results of the above techniques with those resulting from artificial intelligence techniques, we have opted for using *case-based reasoning systems* (CBR)¹, a technology that solves problems on the basis of previous experiences and

¹ Case-based reasoning systems are being widely used to solve business problems. For an insight into this technique see the work of Kolodner (1993).

shows the situations most similar to the actual one. In order to gather the most similar cases to the actual one, in practice two different techniques are available: the so-called “nearest neighbour” and *inductive recovery*. In the nearest neighbour recovery the system selects the case whose attributes are most similar to the actual problem by creating a series of clusters. In the inductive method a decision-making tree is built for each specific case using the entropy concept to measure whether it is an informative attribute.

4. Empirical analysis using discriminant analysis

We have started from our sample made up by 279 firms. Part of the total sample will serve to validate the model; we have thought it would be enough to use 79 observations for such validation and the 200 other observations for estimating the discriminant function.

First, since our objective is to determine the discriminant power of the whole set of variables all of them are simultaneously included in the model. Thus, two discriminant functions will be calculated with the following weights per variable:

$$D1 = -7.233 - 0.343 \cdot V1 - 0.161 \cdot V2 + 0.955 \cdot V3 + 0.812 \cdot V4 - 0.477 \cdot V5 + 0.181 \cdot V6 - 0.007 \cdot V7 - 0.218 \cdot V8 + 0.325 \cdot V9 + 0.294 \cdot V10 - 0.121 \cdot V11 + 0.388 \cdot V12 + 0.381 \cdot V13 + 0.625 \cdot V14 + 0.294 \cdot V15 + 0.827 \cdot V16$$

$$D2 = -0.738 + 0.952 \cdot V1 - 0.223 \cdot V2 + 1.394 \cdot V3 - 1.320 \cdot V4 - 0.476 \cdot V5 + 0.599 \cdot V6 + 0.071 \cdot V7 + 0.265 \cdot V8 - 1.175 \cdot V9 - 0.327 \cdot V10 - 0.284 \cdot V11 + 0.283 \cdot V12 - 0.093 \cdot V13 + 0.042 \cdot V14 + 0.870 \cdot V15 - 0.153 \cdot V16$$

In order to appreciate which is the variable that contributes most in each discriminant function we looked at the structure matrix. According to this matrix the most discriminant variable is firm rating in the first function and economic situation in the second one.

We are interested in knowing which of the selected cases were correctly classified, to achieve this we focused on the comparison between the real group and the predicted group resulting from the analysis, without neglecting the probability associated to being in that group. The results obtained allow us to state that more than half of the errors take place among the credits belonging to the group of loans granted under certain conditions. This is because this group includes loans with quite different characteristics; less than half of the errors corresponded in almost equal proportions to the granted and rejected loan applications. The summary of all cases is as follows: From the selected sample, i.e. the first 200 cases, 79,5% of them have been correctly classified, from the non-selected sample, the other 79 cases, the classification was correct in 70,9% of the cases. It is also possible to know the type of error² that has occurred in them; although there is no large difference between the type of error detected, there is a slight bias to type I error, and therefore it is possible to infer that the system takes more risks than the entity's analyst.

² We have considered two types of errors: Type I Error: The CBR system is takes more risks than the entity's analyst. Type Ii Error: The CBR system is more conservative than the entity's analyst

Table 1. Structure Matrix.

	Function 1	Function 2
Firm rating	0.651*	-0.094
Risk rating	0.644*	-0.107
Positive image	0.474*	0.002
Interview opinion	0.456*	0.146
Management appraisal	0.424*	-0.089
Negative information	-0.413	-0.272
Evol. of economic situation	0.411*	0.024
Evol. of financial situation	0.339*	-0.001
Industry classification	0.289*	0.225
Financial situation	0.211*	-0.165
Double set of books	-0.210	0.131
Economic situation	0.458	-0.542*
Legal regulations	0.315	0.451*
Appropriate fund allocation	0.331	0.339
Property deeds	0.025	-0.266
Net worth statement	-0.40	0.195

Combined intra-group correlations between the discriminant variables and the typified canonical discriminant functions. Variables organised according to size of function correlation.* Greater absolute correlation between each variable and any discriminant function.

The results obtained allow for an evaluation of the effectiveness of the discriminant function; to assess such function other statistics must be taken into account: canonical correlation and autovalues.

Table 2. Canonical Correlation.

Function	Autovalue	% of variance	Cummulative %	Canonical correlation
1	2.129 (a)	91.8	91.8	0.825
2	0.129 (a)	8.2	100.0	0.400

(a) The two first canonical discriminant functions have been used in the analysis

As it can be observed, the first function explains 91,8% of total variability while the second one explains the rest, 8,2%. It is quite normal that the first variable explains most of it and the second one contributes very little information.

5. Empirical study using logistic regression.

To carry out this study we have used a multinomial logistic regression, in which the dependent variable is the analyst's proposal, called decision, and the independent variables are included as covariables. It is a multinomial regression because the dependent variable in the study includes three possible answers. The chi-square classification table and goodness of fit is used. Results are discussed below. The model fit information table shows its signification test. In this case the nil hypothesis states that the model is not significant; since p-value equals zero, we can say that the nil hypothesis is rejected.

Table 3. Model fit information table

Model	- 2 likelihood logarithm	Chi-square	gl	Sig
Intersection only	589.275			
Final	273.085	316.190	32	0.000

In addition to determining that the model is significant, it is necessary to see how do data fit: it must be verified how probable are the results obtained for the estimated model (by maximum likelihood). The nil hypothesis specifies that the model fits data well; in this case, being the p-value high, such nil hypothesis is maintained, and therefore, it can be said that this model fits data well.

Table 4. Data fit

	Chi-square	gl.	Sig.
Pearson	526.779	498	0.180
Desviation	271.463	498	1.000

It is also necessary to analyse each of the variables separately to test whether they are significant. To this the likelihood ratio tests are used; in this case, the nil hypothesis says that the variable is not significant. The table shows the p-value of each variable; if this p-value is less than 0.05 the nil hypothesis is rejected and therefore the corresponding variable is not included in the study. Accordingly, the non-significant variables in this model are the ones labelled as v.2, v.5, v.7, v.8, v.10, v.11 and v.15.

Table 5. Likelihood ratio tests

Effect	- 2 likelihood-log of the reduced model	Chi-square	gl.	Sig.
Intersection	350.678	77.593	2	0.000
V.1	285.685	12.600	2	0.002
V.2	274.134	1.049	2	0.592
V.3	280.467	7.382	2	0.025
V.4	285.059	11.974	2	0.003
V.5	274.032	0.947	2	0.623
V.6	288.271	15.186	2	0.001
V.7	273.251	0.166	2	0.920
V.8	274.787	1.702	2	0.427
V.9	284.725	11.640	2	0.003
V.10	275.907	2.822	2	0.244
V.11	275.637	2.552	2	0.279
V.12	289.808	16.723	2	0.000
V.13	283.176	10.091	2	0.006
V.14	287.997	14.912	2	0.001
V.15	274.290	1.205	2	0.547
V.16	283.811	10.725	2	0.005

The chi-square statistic is the difference in terms of - likelihood-log between the final model and the reduced model. The reduced model is built by deleting an effect of the final model. The nil hypothesis is that all parameters of such effect are 0.

To find out the load of each variable in the model parameter estimates are used. Each category of the dependent variable is compared to the category the system uses by default, in this case, to award the grant. Comparisons are therefore as follows: to reject vs. to grant, and to grant under certain conditions vs. to grant. Quotient B measures change in probability ratio (i.e. change in decision-maker's opinion) when the corresponding explanatory variable increases by one unit.

The Wald test following a chi-square distribution with one degree of freedom will be used to test the nil hypothesis, whether the independent variable influences the probability of the event representing the dependent variable, i.e. whether each of the independent variables do influence the possible granting of the loan. If p-value is greater than 0.05 the variable is not significant to distinguish between the two ranks specified above. For example neither v.2 nor v.6, among others, are appropriate to distinguish between granted loans and rejected loans. However, although variable v.6 is significant to distinguish between granted loans and granted conditional ones, in this case v.2 is not significant either.

The final results obtained with this regression are as follows: from the 63 rejected loans, the system classifies correctly 53 and incorrectly the remaining 10, which means that rejected loans were correctly classified in 84.1% of all cases; of the 90 loans granted under certain conditions, 7 of them were classified as rejected and 33 as granted, i.e. 36.7% of such category was incorrectly classified. Finally, almost all granted loans were correctly classified, since from the 126 cases included in this category 107 were correctly classified 18 were granted under certain conditions and only 1 was rejected. The overall percentage of correct classification is 77.8% and the systems takes more risk than the entity's analyst, like in the case of discriminant analysis.

6. Empirical study using CBR

The inductive method consisting of building a decision-making tree for each specific case was used in order to find out how the CBR technology classifies each of the cases included the sample; its use is more appropriate when the dependent variable can be defined, as it is our case, and when there are several examples of each of them are available, as it is again the case of our sample.

In order to get the most out of data the work was done drawing one case of the sample at each time, which is the one that would serve as test. Thus it was possible to see how the CBR classified each of the loan applications and then match the result with the actual decision of the risk analyst of the banking entity.

From the 279 cases included in the sample under study we detected that 12.18% of all loans were incorrectly classified, which means that 34 of the 279 cases were classified by the system differently from the actual result.

After examining how the system classified the cases included in the sample it is interesting to know which of them were incorrectly classified. However, it is even more important to know which are the most significant variables that have been take into account by the system, because this would let us know which are the fundamental and accounting aspects in the decisions regarding the granting of corporate loans. Although there is not much difference between the percentage of failed cases for each type of error,

type I errors are more frequent: from the 34 incorrectly classified cases, 585 (20 cases) were due to the fact that the CBR system takes more risks than the banking entity.

Furthermore, it is possible to know which are the variables taken into account by the system to classify loans. Since the inductive method presents a decision-making tree for each case, it is possible to know which tree has been used most for granted loans and which for rejected loans; however, it is very difficult to draw a tree model for loans granted under certain conditions, since the corresponding partitions are quite diverse. The most relevant variables drawn from the above model trees are as follows:

1. Economic situation of the firm
2. Firm rating
3. Firm image

7. Statistical techniques vs, artificial intelligence techniques.

There are several studies that have compared results obtained by artificial intelligence technologies and those resulting from the application of different statistical techniques.

In recent years several comparative studies between statistical techniques and the artificial neural nets technology have been carried out in the business world, most of them aimed at predicting corporate insolvency. Hereto, Bell et al. (1990) conducted a study using neural nets and compared its results to those of logistic regression; the conclusions of their study depend on the costs allocated to the different types of errors and evidence the advantages of using nets. Similar to this work is one by Boritz et al. (1995), although this includes two different types of nets. On the other hand, De Miguel et al. (1993) made a comparison between the above-mentioned neural nets and for statistical models (linear probability analysis, logit analysis, probit analysis and discriminant analysis) to predict bank crises; the results from this study are in all cases favourable for neural nets. Furthermore, Martinez de Lejarza (1996), designed a neural net and compared it to discriminant analysis; the results obtained by the net were 96% (approx.) correct, and they have a better performance than those of discriminant analysis, which ranged between 72% and 92% (both approx.). A further paper comparing neural net and discriminant analysis results to study business failure was published by Crespo (2000); these results show a slight advantage of neural nets on the statistical analysis. De la Torre et al. (2001) carried out a similar study, using neural nets and logit analysis for comparison; however, their results did not show any significant differences between both techniques. It is also interesting to underline the work of De Andres (2001) comparing the SEE5 artificial intelligence algorithm to logit analysis and discriminant analysis to make a classification of small and medium-sized firms. This study shows that the SEE5 algorithm outperforms logit analysis but not discriminant analysis.

Comparisons using the CBR technology applied in this study have not been applied yet to the business world. A study by Musgrove and Davies (1995), applied to the field of medicine, compared discriminant analysis and CBR and their conclusion was that CBR was more accurate: the CBR system classified correctly 87% of cases (using the nearest neighbour recovery) as compared to 67% obtained by discriminant analysis. A further study conducted by Nakhaeizadeh (1993), compared both techniques too (using

in this case the inductive method for the CBR) for the prediction of time series achieving similar conclusions in his results.

8. Comparison between discriminant analysis and logistic regression and cbr

The results obtained from the comparison between discriminant and logit analysis and artificial intelligence technologies in the above studies agree to a great extent with those of the empirical study conducted in this work, which uses the CBR artificial intelligence technology. In the empirical analysis reported in this paper there is a clear difference of CBR systems as compared to discriminant analysis and logistic regression. In the case of discriminant analysis the difference between correct classifications by this method and by CBR is 9% approximately; in the case of logistic regression the difference accounts for 8%.

The results obtained from the different technologies show differences in ration of correct classification of loans into one or another category, however, they are similar with regards to the variables considered to be more relevant in the study. Discriminant analysis shows firm rating as the variable with a greater weight in the first discriminant function and the economic situation of the firm as the most relevant variable in the second discriminant function. Logistic regression shows the variables considered to be not significant for the model; this analysis does not exclude the variables firm rating and economic situation, which agrees with the results of discriminant analysis. Other variables pointed out as relevant ones are corporate image, risk rating and personal interview with firm manager. From the CBR model it is possible to know the variables most frequently taken into account in the decision-making trees generated by the system. These variables, as it was mentioned above, are economic situation of the firm, firm rating and the banking entity’s image of the firm.

The results obtained and the most relevant variables in each of the analyses are shown in the following chart (Table 6).

The errors that have occurred in each of the software systems used evidence the differences in loan classification as compared to the classification proposal made by the banking entity analyst. These errors are mostly type I errors, both in discriminant analysis, logistic regression and CBR. Therefore, we can state that all the technologies used have a riskier behaviour than the banking entity analyst.

Table 6. Results obtained

ANALYSIS	ERRORS	VARIABLES
Discriminant analysis	23%	Firm rating, Economic situation
Logistic regression	22%	Economic situation, Firm rating, Client image, Risk rating Personal interview
Induction (CBR)	12,83%	Economic situation, Firm rating Client image, Net worth statement

9. Conclusions

Based on the empirical study carried out it is possible to state that in credit exposure analysis CBR systems are more accurate than the statistical techniques usually applied in this type of decisions.

The percentage of error when using CBR systems (12'83%) allows us to conclude that these intelligent systems are a good support in decision-making for risk analysts from banking entities; however, the task of the banking entity risk analyst should not be neglected; since such decisions are influenced by many external factors we can state that it is impossible to simulate the analyst's behaviour in all situations.

The most important variable for the CBR system is the firm's economic situation, followed by firm rating and by firm image, which agrees with the most relevant variables drawn from discriminant analysis and logistic regression.

References

- Bell, T.; Ribar, G. y Verchio, J (1990): "Neural Nets Versus Logistic Regression: A Comparison of Each Model's Ability to Predict Commercial Bank Failures", *Proceedings of the 1990 Deloitte&Touche /University of Kansas Symposium on Auditing Problems*: 29-53.
- Boritz, J.; Kennedy, D. y Albuquerque, A (1995): "Predicting Corporate Failure Using A Neural Network Approach", *International Journal of Intelligent Systems in Accounting, Finance and Management*, vol. 4, n. 2: 95-112.
- Crespo Domínguez, M. A. (2000): "Una Aproximación a la Predicción del Fracaso Empresarial Mediante Redes Neuronales", *IX Encuentro de Profesores Universitarios de Contabilidad*, Las Palmas de Gran Canaria: 591-607.
- De Andres, J. (2001): "Statistical techniques vs. SEE5 Algorithm. An Application to a Small Business Environment", *The International Journal of Digital Accounting Research*, vol. 2: 157-184.
- De la Torre, J. M.; Merelo, J. J. y Román, I. (2001): "Forecasting Business Failure. A Comparison of Neural Networks and Logistic Regression for the Spanish Companies", *European Accounting Congress*, Atenas, Grecia.
- De Miguel, L. J.; Revilla, E.; Rodriguez, J. M. y Cano, J. M. (1993): "A Comparison Between Statistical and Neural Networks-Based Methods for Predicting Bank Failures", *Proceedings of the III^a International Workshop and Artificial Intelligence in Economics and Management*, agosto, Portland, Oregon.
- Gabás Trigo, F. (1997): "Predicción de la insolvencia empresarial", *Predicción de la insolvencia empresarial*, Monografías de AECA: 13-31.
- Kolodner, J. L. (1993): *Case-Based Reasoning*, Morgan Kaufmann Publishers, Inc., San Mateo, CA.
- Mar-Molinero, C. y Serrano-Cinca, C. (2001): "Bank Failure: A Multidimensional Scaling Approach", *The European Journal Of Finance*, vol. 7:165 –183.
- Martínez de Lejarja (1996): "Forecasting Company Failure: Neural Approach versus Discriminant Analysis. An Application to Spanish insurance companies", *Intelligent Systems in Accounting and Finance*, Sierra y Bonsón (eds.): 169-185.
- Mora Enguános, A. (1995): "Utilidad de los Modelos de Predicción de la Crisis Empresarial", *Revista Española de Financiación y Contabilidad*, vol. XXIV, n. 83, abril-junio:281-300
- Musgrove, P. B. y Davies, J. (1995): "A Comparative Study of Three Machine Learning Approaches to the Treatment of Patients of Anticoagulant Out-Patient Clinics", *Applications and Innovations in Expert Systems III*, Macintosh and Cooper (eds.), Oxford.
- Nakhaeizadeih, G. (1993): "Learning Prediction of Time Series. A Theoretical and Empirical Comparison of CBR with some other Approaches", In *Proceedings of EWCBR'93*, Richter et al. (ed.), Berlin.
- Núñez García, C. (1997): *La Construcción de una Red Neuronal para el Análisis de Riesgos en las Entidades Financieras*. Tesis doctoral, Universidad de Sevilla.