QÜESTIIÓ, Vol 18, 3 pp. 385-395, 1994

A MODEL FOR CREDIT SCORING: AN APPLICATION OF DISCRIMINANT ANALYSIS

MANUEL ARTÍS, MONTSERRAT GUILLÉN and JOSÉ Mª MARTÍNEZ*

Universitat de Barcelona

The application of statistical techniques in decision making, and more specifically for classification requirements, has proved to be adequate in the context of financial problems. In this study, we present the methodology used and the results obtained in the elaboration of a decision-support system for credit assignment. The problem was to provide an automatic tool for a Spanish financial institution that needed to quantify and analyse credit applications from clients. Firstly, we shall present the statistical techniques. Secondly, we shall describe the characteristics of the data set used for estimating the discrimination function and, finally, we show the results obtained when the model is used to discriminate among clients in the data set, whose history of financial behaviour is reflected in the data by means of a variable counting the number of unpaid instalments. Essentially, every individual asking the bank for a loan is assigned a certain score. This score is directly related to the probability he or she has of returning the money, but also to the risk of the institution lending that amount. Some comments about the application of the model are given and results concerning the optimal level of risk are also discussed, in order to give clear patterns for implementation.

Key words: Multivariate Analysis, Discriminant functions, Classification, Credit Scoring.

---

* Manuel Artís, Montserrat Guillén and José Mª Martínez. Departament d'Econometria. Estadística i Economia Espanyola. Universitat de Barcelona. Avgda. Diagonal, 690. 08034 Barcelona. Spain.

We acknowledge the support received from DGICYT, grant PB92-0545.

-Article rebut el novembre de 1993.

1. INTRODUCTION

The application of statistical techniques for the analysis of decision problems entailing an element of classification has experimented a great development in the last decade. Specially, those techniques have proven to be very useful in the context of business and finance (see Altman, Eisenbeis and Sinkey, 1981).

This work is a real application of the elaboration of a model for decision making, based on discriminant analysis, that has been developed to provide automatic decisions for credit assigment to private clients in a finance institution in the Spanish territory. More explicitely, we present a methodological approach that leads to the elaboration of a model. Afterwards, we present the characteristics of the data base used for the estimation and, finally, we analyze the results obtained for the simulation of the model performance as a tool for the evaluation of the risk adquired by the institution in the acceptation of loans.

The main interest of this paper is to present a practical application of the discriminant analysis techniques and, also, in the original use of the data in the context of a Spanish bank.

2. METHODOLOGY

Discriminant analysis concerns two different types of techniques, Factorial Discriminant Analysis and Discriminants Functions. The first one is considered in the analysis of differences among groups of individuals, it reveals the characteristics that make their relevance clear and also shows their relative importance. Discriminant functions have the aim of finding an optimal way of classifying units in groups. This latter problem has been approached in many different ways.

One of the applications of discriminant analysis is we studied in this paper, and it deals with the classification of individuals applying for a credit into two groups depending on whether they will pay the credit back on not. This procedure is based on assigning a probability to every individual, it is the probability of paying the money back or otherway, the probability of not paying it back. The probability may easily be transformed into a score that will be the essential information to make the final decision about the acceptation of the application for credit. Somehow, the probability of not returning the money is a measure of the risk of the financial institution when the client will be given credit, and

it may serve as a necessary information for the global estimation of risk of the financial institution in a given period of time.

In such a situation, a data base is normally available, containing information about clients that have applied for credit, and whose behaviour is known throughtout the period of credit payment. Moreover, usually, a batch of characteristics of the individuals, private characteristics, socio-economic status and financial behaviour in the past are available so that a typical portrait can be established for the analysis of future applications. Let us remark that the sample is truncated. This means that it has been chosen in the subpopulation of customers that already have credit, and it could be non-representative of the larger population of people that have applied for it. Nevertheless that was the only available information in the data bases and that limitation was not relevant for the practical aims of our study.

Let us call $\mathbf{x}$ the vector of the values for the explanatory variables measured in the individual whose application is to be evaluated. Let us assume that $\mathbf{x_M}$ is a vector of averages for the explanatory variables for the group of sampled individuals that have effectively payed the credit back, and $\mathbf{x_{NM}}$ the corresponding vector for the group that have not payed the money. Let us denote by $\mathbf{S_M}$ the covariance matrix for the first group and $\mathbf{S_{NM}}$ for the second. Afterwards, we may establish a measure that calculates the distance from individual to group that will be different for every group. The idea is that an individual will be classified into the group that is nearest to him, using the following distance:

$$D^2_M(x) = g_1(\mathbf{x}, M) + g_2(M)$$

for the individuals paying back,

$$D^2_{NM}(x) = g_1(\mathbf{x}, NM) + g_2(NM)$$

for the individuals not paying back.

The form of the functions is as follows:

$$g_1(\mathbf{x}, i) = (\mathbf{x} - \mathbf{x_i})^T \mathbf{S_i}^{-1} (\mathbf{x} - \mathbf{x_i}) + \log |\mathbf{S_i}| \quad i = M, NM$$

and,

$$g_2(i) = -2 \log(q_i) \quad i = M, NM$$

where q_i is the probability *a priori* for every group.

Sometimes, it is possible to simplify the above mentioned formula, due to the fact that the *a priori* probabilities may be consedered equal. Another source of simplification is the use of the global variance and covariance matrix, calculated

for all the individuals in the sample, that will be used instead of the two matrices above.

This latter aspect is a common practice when there are small differences in the dispersion and correlation between the variables in the studied groups. The specialised literature states that, if a test is performed, it may suggest to use the global matrix, but it is necessary to use the hypothesis of normality to ensure the adequate statistical properties of the estimators. Finally, the probability of belonging to each group is given by the Bayes' rule and it will provide the scoring:

$$p(\mathbf{x}) = \frac{\exp(-\frac{1}{2}D^2_{NM}(\mathbf{x}))}{\exp(-\frac{1}{2}D^2_{NM}(\mathbf{x})) + \exp(-\frac{1}{2}D^2_{M}(\mathbf{x}))}$$

From the early works of Fisher (1936), a great deal of studies have been published about Discriminant Analysis Applications, the classification problems have been viewed from very different points of view, including non parametric techniques.

3. THE DATA

The elaboration of a data base has the objective of obtaining reliable and detailed informationo about a number of clients that have been given credit, so that they would be represented randomly in Spain. Initially, a sample size of 5000 individuals was suggested, the information of their characteristics when the application was done had to be complemented with the incidences at payment time.

With a considerable effort, due to the difficulties in the data collection process, it was possible to obtain 4961 individuals, and the variables of interest were the year of birth, gender, marital status, average anual income, account level, studies, number of children, property of the household, job, credit destination, ... We quantified the qualitative variables although some aspects might depend on that quantification.

Afterwards, a simple statistical analysis was performed for a first analysis and detection of errors was carried out. Next, we proceeded to the model selection.

4. FINAL MODEL FOR CREDIT SCORING AND EVALUATION OF ITS DISCRIMINANT PERFORMANCE

In the first step, all the variables were introduced in the model so that a backstep algorithm might be used to find out those with greater capacity for discrimination. This search was done step by step. The objective was to obtain the best model, that is to say the model which provided the best proportions of correct classification when applied to the individuals in the sample used for estimation. We will consider a good model for credit scoring, a model that, when evaluating the applications in the sample leads to a greater percentage of correct classification of the individuals according to their posterior behaviour concerning the return of the amount. That means, finding out whether the applications belong to the group of paying or not paying back. This statistical procedure will lead to a model that is not necessarily the optimal, but it is the best when compared to the different trials. The search for a good model depends on the number of variables that the user allows into the model, or it depends on the restrictions about the presence or absence of certain variables. The first difficulty appears when deciding the maximum number of variables to be present in the model, e.g. the degrees of freedom.

Finally, after the different approximations and models, the percentages of correct classification for both the entire sample and the two populations were calculated. Simultaneously, some sensibility studies were performed in order to decide the inclusion or exclusion of some variables, in fact, we were trying to see the changes produced by the elimination of some variables that are difficult to measure in front of the inclusion of much more accessible variables.

The variables included in the final model are divided into different groups according to the source of the information that they provided, whether they are items responded by the individual applying for credit, or if those items are information that although it may be provided by the applicant, the financial institution is able to check in its archives.

The variables in the model may be found in one of the following three groups:

- Personal variables —three— (date of birth, marital status, number of children,...)
- Socio-economic variables —four— (net monthly income, ownership of house,...)
- Financial variables —five— (Monthly mortgage, availability of credit card,...)

The main innovation about the variables that are used in the model for credit scoring is that the above variables provide the information that is needed to create new variables finnally used in the model. The modifications are made in two different senses and have two well established objectives:

a) On the one hand, the financial institution is generally interested in preserving the confidentiality of the discriminant function that is finally being implemented. Since the scoring for a particular applicant must be calculated using special purpose software, it might well be the case that someone could have access to the coefficients for the discriminant function. We have seen that many credit managers sometimes distort the information in order to obtain a certain value of the scoring because they know the influence of a variable in the final result. The construction of new variables tries to avoid the above mentioned practice, or at least make it more difficult.

b) On the other hand, by using some new variables, the model can cope with the interactions which are produced among the different variables. Therefore, some new variables, interactions, have been introduced to detect very specific groups of defaulters. The interaction variables lead to a much better performance of the final model.

Once the variables that are to be present in the discriminant function are decided, we may proceed to the evaluation of the discriminating power of the model, or the capacity and operativity of the automatic decision tool. So, the discriminant function is used for the classification of the sampled applications, which are known to have complete information and known posterior behaviour, since the bank has in the files the number of payments that they have not been paid. Table I. shows five different columns, every row corresponds to a different prior probability for the two populations, this is equivalent to say that each row corresponds to a different threshold value for the posterior probability . The threshold establishes the minimum value of the scoring for the individual to be classified into the group of non defaulters, and for the credit to be granted.

Starting with a value of 50, and until 80 points (note that the posterior probability has been multiplied by 100), this table shows the different sceneries. Column one contains the different thresholds. The values in columns two and three correspond to the credits granted, and are the number of good credits granted and the number of bad credits granted, respectively. The next two columns correspond to the denied credits, first the number of good credits denied and finally bad credits denied.

Some graphics are presented below, showing a picture of the information in Table 1. Figure 1. shows that when the threshold increases, that is the minimum

score for granting, then the percentage of applications granted decreases, as well as the number of bad applications accepted.

Table I

THRESHOLD	GRANTED		DENIED	
	GOOD	BAD	GOOD	BAD
50	2486	186	1343	676
51	2470	183	1359	679
52	2449	177	1380	685
53	2429	174	1400	688
54	2411	170	1418	692
55	2388	165	1441	697
56	2365	160	1464	702
57	2343	154	1486	708
58	2329	151	1500	711
59	2306	150	1523	712
60	2290	149	1539	713
61	2257	147	1572	715
62	2232	141	1597	721
63	2212	134	1617	728
64	2180	129	1649	733
65	2149	125	1680	737
66	2104	115	1725	747
67	2062	108	1767	754
68	2026	106	1803	756
69	1956	94	1873	768
70	1866	80	1963	782
71	1799	79	2030	783
72	1707	71	2122	791
73	1608	66	2221	796
74	1534	65	2295	797
75	1441	57	2388	805
76	1319	49	2510	813
77	1231	44	2598	818
78	1097	34	2732	828
79	987	29	2842	833
80	897	22	2932	840

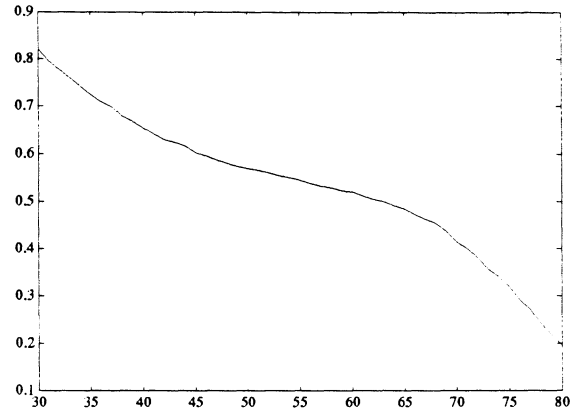

Figure 1

Percentage of granted credits for differents minimum score levels.

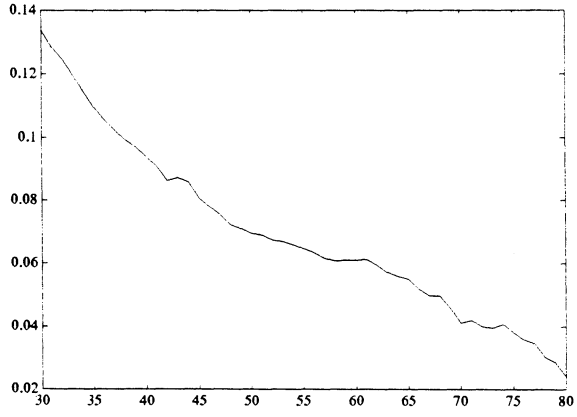

Figure 2

Percentage of bad applicants among those granted.

Looking at Figure 2, we see that it is obvious that between threshold 60 and 70, there is a zone of fluctuations. Therefore we suggested that applications receiving a score between those two levels should be regarded as doubtful, and should be analysed more carefully.

Let us make an example, if we decide that 80 will be the minimum score for granting credit, then only 919 credits would be granted of the inicial 4691 in the sample. 22 bad credits and 897 good credits. On the other hand 3772 would be denied , from those 2932 would be good and 840 bad. If we believe that this minimum score is too high then we can set the threshold to 79 points,

with this new minimum, 97 more applications are accepted and of those only 7 are bad credits. Using this strategy, we establish the optimal threshold. The best minimum point is about 60, according to the previous criterium of the financial institution based on profitability and risk. Therefore, the bad group is correctly classified (denied) with a percentage of 83%. 61% of the good clients are correctly classified (accepted). 52.36% of the applications would be accepted. Among those accepted only 6.1% would be bad credits, which means about 1 in every 20 credits would be delinquent.

5. STATISTICAL PROCEDURES AND SOFTWARE FOR THE AUTOMATIC EVALUATION

5.1. Discriminant function estimation

All the statistical procedures applied used the sample of almost 5000 applications and were performed using the computer statistical analysis system SAS. At the begining the basic procedures were used, for data description and depuration. For the elaboration of the decison model, we used STEPDISC and DISCRIM, that allow the estimation of the discrimant function of the model. Finally, we developed a tool in SAS to evaluate the results of the estimated models. Once, we found the final model, the following step was to develop a special purpose software to automatize the evaluation of new applications according to the model found.

5.2. Software for the automatic evaluation of new credit applications

For the practical implementation, we have designed a new autonomous system for the evaluation of applications on the basis of the classification model selected. Moreover, with the use of this software, the calculation of the final score remains as a black box, not the credit manager nor the applicant know the specific weights of the different items of information and their contribution to the final result. Thus, manipulations would be reduced and biased sores would be avoided.

The computer software elaborated by the authors has the objective of implementing the final model. The system has been prepared for IBM compatibles under DOS. It has been written in the programming language FORTRAN. The

software has two different modules. Firstly, the interactive module reads the items of information required by asking the applicant or the credit manager. Secondly, the scoring module uses the information to create new variables and calculates the final score that will be between 0 and 100, according to the posterior probability of the applicant to pay back.

A remarkable characteristic of the system is the facility to adjust for new estimations of the model that may vary along time.

6. FINAL REMARKS

The evaluation of credit applications, for loans of moderate quantities, may be supported by a quantitative tool based on the historical behaviour of applicants that were granted credit.

The main characteristic of the selected discriminant model is the powerful evaluating facility, the simplicity of the information required to give the final score. The applicant will normally inform only about a limited number of information items, which the bank must verify to ensure its reliability.

The statistical methodology applied to the context of finantial problems has lead to the development of the decision tool. This entails an innovation in speed and computerisation of the problem of granting credit, that will undoubtedly contribute to provide a dinamicity specially adequate for this kind of transactions, faster and providing better service to the client.

The last conclusion has to do with the validation of the presented model for classification. On the one hand, we must take into account that there must be a continuous reviewing of the evolution of the portfolio, analysing the behaviour and its changes. The percentages of defaulters should be frequently supervised to see whether the politics of granting must be altered so that the global risk acquired by the institution is not increased.

BIBLIOGRAPHY

[1] **Anderson , T.W.** (1958). *An Introduction to Multivariate Statistical Methods.* John Wiley and Sons.

[2] **Altman E.R., Avery R., Eisenbeis R. and J. Sinkey** (1981). *Classification Techniques with Applications to Business and Finance.* JAI Press.

[3] **Boyes, W.J. Hoffman, D.L. and S.A. Low** (1989). "An Econometric Analysis of the Bank Credit Scoring Problem". *Journal of Econometrics*, **40**, 3–14.

[4] **Fisher, R. A.** (1936). "The use of multiple measurements in taxonomic problems". *Ann. Eugen.*, **7**, 179–188.

[5] **Hand, D.J.** (1981). *Discrimination and Classification.* John Wiley and Sons.

[6] SAS User's Guide (1983). Sas Institue, Cary , U.S.