Pattern Recognition in Credit Scoring Analysis*

Maria Teresinha Arns Steiner *Celso Carnieri*

Departamento de Matemática
Universidade Federal do Paraná
Curitiba - Paraná - Brasil
Caixa Postal 19.081 - CEP 81531-990, Fax: (041)267-4236
tere@mat.ufpr.br, carnieri@mat.ufpr.br

Abstract

Recognizing and foreseeing which credit clients will be "good or bad payers" is an important and difficult task for bank institutions and credit protection services. Using data from approximately 10,000 clients obtained from a large private Brazilian bank, we present a methodology to perform the credit scoring analysis. The methodology proposed is divided into 2 stages: statistical data analysis and the use of a model to perform the Pattern Recognition, discriminating the two groups mentioned earlier.

Keywords: Pattern Recognition, Credit Scoring, Multivariate Analysis.

1 Introduction

Credit operations such as loans, funding, guarantees, credit cards, or bank overdrafts make up a substantial part of commercial bank revenues. Any mistake in the credit decision-making means that a single operation may cause the loss of the profit obtained from tens of successful others. Correct credit decisions are essential for the survival of banking enterprises. Therefore, it is desirable and necessary to analyze a business proposal and compare the "cost of granting" versus "the cost of denying".

The risk / return relationship is implicit in credit granting. The volume of bad debts, as well as their profitability, are results of the policy adopted by the organization and its criteria of credit granting. The optimization of results is, therefore, a consequence of an efficient credit policy, related to the collection policy and other company policies [8].

*The first author was supported by CNPq.

The credit policy of a commercial bank comprises the establishment of interest rates, terms of payment, guarantees and risk level related to each kind of operation. Another relevant factor in the credit policy of a commercial bank is the government economic policy, implemented via Central Bank of Brazil, in this case.

The analysis of the decision-making process is highly complex, because it involves not only the previous experience of the bank analysts, but also tools and techniques that may help them in this task.

Operations Research methods, widely used in this type of analysis, take into account historical records to decide whether to grant credit. Such techniques, among which we may quote Linear Programming and Logistic Regression Models, are efficient tools to assist credit managers, if correctly employed.

The advantages of employing Operations Research techniques in credit management are, among others [7],[2]:

- more creditworthy people will be granted credit (or additional credit), increasing profits;

- more non-creditworthy people will have their credits denied (or reduced), thus decreasing losses;

- credit applications may be rapidly processed;

- decision-making is objective and does not go through subjective criteria;

- less people are needed in credit management, and more experienced people may concentrate on the most difficult cases.

This paper focuses on the Logistic Regression statistical model that has shown itself promising in Pattern Recognition analysis. The approach of this technique and its application to the real problem, described in section 2, is found in section 4. Prior to the application of this technique, a statistical analysis is performed on the data in section 3 and some sub-sections of section 4. Conclusions are presented in section 5.

2 Description of the Real Problem

To effectively analyze the problem to be approached, a data survey from a private commercial bank in Brazil was performed. The application data of 9,942 customers were surveyed, of which 6,658 had been classified by the bank as "creditworthy" or current, and 3,284 "non-creditworthy" or defaulting.

The data survey on each one of those customers, totaling 25, are disclosed in Table 2.1; part of the data is divided into categories. Part of the data, as may be observed,

had already been quantified by the bank. Another part, however, had to undergo certain adjustments to be quantified: sex (male - 1, female - 0); date of birth (the age was obtained), municipality (capital - 1, countryside - 0); activity code (taken from a list of the Internal Revenue Service, the codes were transformed into values ranging from 1 to 582), period of employment (calculated in years), position (taken from the same Internal Revenue Service list, the codes were transformed into values ranging from 1 to 582).

The following individual customer characteristics were disregarded: District - it is an interesting information, but since the data are nation-wide, it is a practical impossibility to quantify that value. For the same reason, the following characteristics were also disregarded: Postal code State, District of work, Postal code, City of work, State of work.

On the other hand, the characteristic Region was added with the categories North, Northeast, Southeast, South and Center-West with the values 1, 2, 3, 4 and 5, respectively, as shown in Table 2.1. Thus, 19 characteristics (variables from A to S in Table 2.1) were obtained.

Furthermore, out of the 9,942 customers, several were discarded because their application data were incomplete or had problems: information deemed important were blank or filled with absurd figures, evidencing mistakes in their filling. Therefore, after discarding all those, a sample of 6,470 customers was obtained, 4,361 current and 2,109 defaulting.

Variable	Data Customer	Categories	Quantitative Data
1-A	Sex	Female	0
		Male	1
2-B	Marital Status	Single	1
		Married	2
		Widow	3
		Divorced	4
		Separated	5
		Others	6
3-C	Date of Birth	—	(age)
4-D	Nationality	Brazilian	1
		Foreign	2
5-E	Education	Elementary	1
		High School	2
		College	3
		Others	4
6-F	Current Residence	—	(years)
7-G	Former Residence	—	(years)
8-H	Residence Type	Own	1
		Rented	2
		Functional	3
		Others	4
9-I	Municipality	Interior	0
		Capital	1
J	Region	North	1
		Northeast	2
		Southeast	3
		South	4
		Center-West	5
10	District	—	(name)

(to be continued)

11	Postal Code	—	(number)
12	State	—	(abbr.)
13-K	Activity Code	(582 codes)	(code-number)
14-L	Income	—	(amount)
15-M	Term of Employment	—	(days)
16	District of Work	—	(name)
17	Postal Code of Work	—	(number)
18	City of Work	—	(name)
19	State of Work	—	(abbr.)
20-N	Position	(582 codes)	(code-Number)
21-O	Spouse income	—	(amount)
22-P	Dependents	—	(number)
23-Q	Properties	—	(quantity)
24-R	Chattels	—	(quantity)
25-S	Insurances	—	(quantity)

Table 2.1 Customers Data
Variables from 1 to 25: given by the bank;
Variables from A to S: considered in this paper.

3 Statistical Data Analysis

The use of statistical techniques may be very useful in the preliminary data analysis, refining the information to be passed on to the Pattern Recognition technique, hence minimizing the time and effort required for their development, increasing the precision of the final model [5],[4]. For those reasons, the techniques presented under 3.1, 3.2, 4.1 and 4.2 were applied.

3.1 Hotelling's T^2 Test

Hotelling 's T^2 test [5] is used to test the equality of the mean vectors in two multivariate populations. Assuming that π_1 and π_2 populations are multivariate with means μ_1 and μ_2, respectively, the hypothesis tested is H_0: $\mu_1 = \mu_2$ versus the alternative H_1: $\mu_1 \neq \mu_2$, which means checking if populations π_1 and π_2 are centered on the same point. If H_0 is rejected, one concludes that the separation between the two populations, represented by their samples U and V, with m and k points and averages $\bar{x}_U$ and $\bar{x}_V$, respectively, is significant, that is, the populations are different in their many average characteristics (or variables) n [5].

Hotelling 's T^2 test may be conducted as follows:

$$(3.1) \qquad T^2 = (\bar{x}_U - \bar{x}_V)'[\left(\frac{1}{m} + \frac{1}{k}\right) S_p]^{-1} (\bar{x}_U - \bar{x}_V)$$

$$\text{with} \qquad S_p = \frac{(m-1)S_U + (k-1)S_V}{m+k-2}$$

where S_U and S_V are the estimators of the variances of the π_1 and π_2 populations based on samples U and V, respectively.

The statistics of the test has an F distribution [6] with $\nu_1 = n$ and $\nu_2 = m + k - n - 1$ degrees of freedom, that is:

$$T^2 \sim \frac{(m+k-2)\,n}{(m+k-n-1)} F_{n,m+k-n-1} \tag{3.2}$$

and defining $T^2L = \frac{m+k-n-1}{(m+k-2)n}(\bar{x}_U - \bar{x}_V)'[\left(\frac{1}{m} + \frac{1}{k}\right) S_p]^{-1} (\bar{x}_U - \bar{x}_V)$

we note that $T^2L \sim F_{n,m+k-n-1}$.

Taking into account only the data concerning customers with a monthly income of R$ 1,000.00 or less, we have a total of 1,717 customers, with m = 1,460 (current), k = 257 (defaulting) and n = 19 (number of variables), and, for this case, Hotelling's T^2 test provided us with $T^2L = 5.82$.

Knowing that $F_{19,1697} = 1.88$ [6] (1% of error), we have $T^2L \geq F_{n,m+k-n-1}$ (that is, $5.82 \geq 1.88$), and it may be concluded that the two populations are different, with a 1% probability of error.

Since in this test the number of current customers is much greater than the number of defaulting ones, a sample of the first set was chosen, remaining 514 customers (m=257, k=257 and n=19), and on that figure Hotelling's T^2 test was once again applied. In this case, $T^2L = 4.38$, and again $F_{19,494} = 1.88$ [6] (1% of error), and the conclusion is that the two populations are different as in the previous result, as expected.

3.2 Correlation Between Data

Considering the set of data presented under 3.1, where m = 257, k = 257 and n = 19, the correlation matrix between the variables was calculated to verify the relationship between them. From the results, we observed that the variables have correlation values smaller than 0.3 determining the independence among them. The few exceptions are the weak correlation between age and term of employment, age and property, and property and chattels variables, where the correlation values are 0.459, 0.332 and 0.435, respectively.

4 The Logistic Regression Model Approach

In Steiner [9], 1995, six Pattern Recognition methods were studied. Of those, two involved the Linear Programming technique, three are statistical, and the last method involves neural networks. Among them, the Logistic Regression method had better performance. For this reason, that method was applied in this work and it is detailed

below:

Logistic Regression, within statistical analysis, consists in relating through a model the response variable with the factors influencing the occurrence of a given event. In this study the aim is to quantify the influence of certain variables, such as a customers assets, on the non-defaulting behavior.

When the random response variable Y, for which an adjustment model is desired, is of the dicotomic type ("1" or "0" responses) and the idea is to study the relation between Y and the other variables (x_1, x_2, ..., x_n), which usually represent characteristics of interest, what is done is to estimate Y using the sigmoid (logistic) mathematical function [3]:

$$Y = f(\underline{x}) = (1 + e^{-\eta})^{-1}, \underline{x} \in R^n \tag{4.1}$$

with $\eta = g(\underline{x})$ obtained by linear adjustment. The quality of the adjustment is measured by the deviation function, defined by [3]:

$$D = -2 \{L_p - L_{(m+k)}\} \tag{4.2}$$

where L_p and $L_{(m+k)}$ are the maximum of the log-likelihood function for the model under investigation with p parameters and for the satured model, respectively.

Considering all the 514 customers in section 3.1, a Logistic Regression model was obtained for the problem with the aid of the GLIM [1] (Generalized Linear Interactive Modelling) computer package, with the consequent classification of each customer and the determination of the main variables, that is, those characteristics that have greater influence on the creditworthiness of a given customer.

In this case, the most important variables were shown to be: A, D, E, H, J and S, that is, sex, nationality, education, residence type, region and insurances. Factor N, position, seems to have no influence at all. In this situation, the percentage of errors in the customer classification, i.e., current customers classified as defaulting and vice-versa was $155/514 = 30.15\%$, a figure considered relatively high. We have to observe that the points (customers) sets U and V used in the model adjustment (or training) were the same used to test the model.

4.1 Factorial Analysis and Variables Transformation

In order to improve the point classification performance, the variables were transformed in an attempt to differentiate as much as possible customers belonging to sets U and V. Since the prognosis "1" or "0" with the least possible error is desired, the relationship of the response variable with the 19 original variables and others derived from those were analyzed on the basis of the deviation function (value provided by GLIM when adjusting the model). Depending on the value of the deviation function

being statistically significant or not, the variable was incorporated to the model or not.

When that transformation was applied to the variables of the 514 customers analysing the deviation function, 27 variables were obtained, and the ones with greater influence on this model were: P, Q, A, lnC, lnH, lnK, E, E^2, H, H^2, C, S, J, J^2, D, all contained in the previous model. In addition to the variables sex, nationality, education, residence type, region and insurance, the following variables were also considered as important: dependents, properties, age and activity code. For this situation the total percentage of errors in the classification was 144/514 = 28%.

4.2 Possible Discarding of Data

To further improve the model performance, customers whose residues (also provided by GLIM) were equal or above 2 were removed from the previous model, since those customers were considered atypical. From the set of data in section 4.1, 20 customers were discarded, 10 from set U and 10 from V, approximately 4% (20/514) of the total. This consideration was defined after discussing with experts in this field. Coincidentally, the number of points discarded was the same for both sets.

Thus we had m = 247, k = 247 and n = 27. With this data, the Logistic Regression Model was again adjusted. For that, 42 customers from U and 70 customers from V were classified wrongly. The percentage of errors in this case was, therefore, 112/494 = 22.67%.

4.3 Analysis of the Results

If, instead of classifying customers directly by the results obtained in the interval [0 ; 1], that is: response in [0 ; 0.5), the customer is defaulting; response in [0.5 ; 1], the customer is current, we chose to subdivide the interval into three parts: response in [0 ; 0,35), the customer is defaulting with automatic classification; response in [0.35 ; 0.65], the customer is doubtful and manually classified by the manager, and response in (0.65 ; 1], the customer is current with automatic classification, we would have the results presented in Table 4.1.

In this case, out of the 494 customers, (150+161)/494 = 62.96% customers have been correctly classified; (65+52)/494 = 23.68% needed manual classification, because they were in the range of doubtful customers, and (21+45)/494 = 13.36% were wrongly classified. We thus have the percentage of errors decreased from 22.67% to 13.36%.

Classification

Responses	Current	Defaulting
(0.65 ; 1]	161/247	45/247
[0.35 ; 0.65]	65/247	52/247
[0 ; 0.35)	21/247	150/247

Table 4.1. Customers classification considering the response interval [0 ; 1] divided into three parts.

5 Conclusions

Aiming to perform Pattern Recognition of current and defaulting customers, the Logistic Regression model - considered one of the best, among the methods researched by Steiner [9], 1995 - was employed. A real set of data, containing application data obtained from a private Brazilian bank, was employed.

Here, the data was modeled for customers with income equal to or below R$ 1,000.00. The performance of the model was considered reasonable, presenting approximately 20% of errors (section 4.2). This relatively high percentage may be explained by the lack of behavioral information on each customer and by the wide variability existing in application of data information.

For this reason, the alternative of dividing the response interval [0 ; 1] into three parts was proposed, automatically classifying customers as defaulting or current, for the responses in the extremes of that interval; and manually classifying doubtful customers, with responses in the center of the interval. In this way the percentage of errors was approximately 13% (section 4.3), evidencing, therefore, a very significant reduction.

References

[1] AITKIN, M., ANDERSON, D., FRANCIS, B., and HINDE, J. Statistical Modelling in GLIM. Oxford, Claredon Press, 1989.

[2] CURNOW, G., KOCHMAN, G., MEESTER, S., SARKAR, D. e WILTON, K. Automating Credit and Collections Decisions at AT&T Capital Corporation, Interfaces, 27, n. 1, 1997, p. 29-52.

[3] DOBSON, A. J. Introduction to Statistical Modelling. New York, Chapman and Hall, 1983.

[4] GORNI, A. A. Redes Neurais Artificiais - Uma abordagem revolucionária em Inteligência Artificial. Micro Sistemas, São Paulo, 1993.

[5] JOHNSON, R. A. and WICHERN, D. W. Applied Multivariate Statistical Analysis. New Jersey, Prentice-Hall, inc., 1988.

[6] MENDENHALL, W., SCHEAFFER, R. L. and WACKERLY,D. D. Mathematical Statistics with Applications. Boston, Massachusetts, Duxbury Press, 1981.

[7] ROSENBERG, E. e GLEIT, A. Quantitative Methods in Credit Management : a Survey, Operations Research, 42, n.4, 1994, p. 589-613.

[8] SILVA, J. P. Análise e Decisão de Crédito. São Paulo, Ed. Atlas S. A., 1993.

[9] STEINER, M. T. A. Uma Metodologia para o Reconhecimento de Padrões Multivariados com Resposta Dicotômica, Tese de Doutorado em Eng. Produção, UFSC, 1995.