The use of graphical models in consumer credit scoring

Elena Stanghellini
Dipartimento di Scienze Statistiche, Università di Perugia
Via Pascoli, 1 – C.P. 1315 Succ. 1
06100 – Perugia, Italy
E-mail: stanghel@stat.unipg.it

1. Models for consumer credit scoring

With the term "consumer credit" we mean "any of the many forms of commerce under which an individual obtains money or goods or services on condition of a promise to repay the money or to pay for the goods or services, along with a fee (the interest), at some specific future date or dates" (Lewis, 1994, p. 1).

Consumer credit has become an enormous business in industrialized countries. It is estimated that in Italy in 1997, the total amount of consumer credits has increased by about 21% over the previous year. This growth is in line with a positive trend started in the second half of the 90's.

The need to cope with a vast demand for credits forced the credit grantors to implement automatic techniques for deciding whether to grant an application or not. Many statistical models, such as neural networks and logistic regressions, have been used for discriminating between good and bad clients, or *risk* (for a review see Hand and Henley, 1997). A crucial role is played by the definition of a good risk and by the choice of the predicting variables.

These models are mainly univariate. In recent years, under the pressure of competition, finance agencies have started to develop new products. The aim is not only to widen their portfolio, but also to keep active relationships with good clients already taken on file and to prevent bad clients from becoming a loss for the agency. As an example of the first situation, it is more and more frequent that clients with already one open relationship with the credit grantors (for instance, a credit card) are offered a second loan with different schemes of repayment. As an example of the second situation, finance agencies have started to sell their own insurance on the loan. A bad client may still be profitable, provided that the client buys insurance on the loan and the defaults occur sufficiently late.

As a result, discriminating between good and bad risks is becoming a process which involves many related outcome variables, which altogether give a measure of whether the clients revealed to be profitable or unprofitable in the whole process. This paper aims to show the potential of graphical models in the described context.

Graphical models are multivariate systems to study associations and dependencies among variables. Each model is represented by a graph where a node is associated to each variable (see Figure 1). Some pairs of nodes are joined by edges, which can be undirected, showing an association, or directed, showing a dependence. The key idea is that of conditional independence between a pair of variables, which is visualized in the graph by the absence of an edge between the corresponding nodes. In the literature, graphs have been used to represent different, though related, models (see Cox and Wermuth, 1996). The relationships between graphical models and other well-known multivariate statistical models are summarized in Wermuth (1992) and Cox and Wermuth (1993). Here we will make use of chain graph models (see Lauritzen, 1996). Other application of these models to credit scoring are in Hand *et al.* (1997), Sewart and Whittaker (1998) and Stanghellini *et al.* (1999).

Table 1: Categories of the variables in the chain graph.

Age	No. Children	Residential Code	Employment Code	Average Funds (Euro)
18–29	0	Owner	Low Income Employee	0–258.2
30–39	1	Mortgage Payer or Renter	High Income Employee	258.3–516.5
40–49	≥ 2	Other	Self Employee	516.6–774.8
50–59			Other	≥ 774.9
≥ 60				

2. The determinants of the behaviour with the credit card

We analyse the behaviour of clients holding a credit card issued by a major italian finance agency. To gain some insights, we first look at a data set formed by the people who started their business in the first six months of 1997 and have an active relationship with the agency at July 1997 (about 30,000). The credit is so called "revolving": for each client the amount of a new credit is summed over the previous ones. The monthly repayment is a fixed proportion of the total credits, which should not be higher than a threshold, plus the interest. Each month the client is classified as "inactive" if the account has had no outstanding balance for more than six months, "dormant" if the account has had no outstanding balance for less than six months, or "active" if the account has outstanding balance.

There is a substantial proportion (about 46%) of these clients who after one year (the last measure available is 98/08) become inactive. About 35% are "dormant" at 97/12 (the behaviour around Christmas is an important variable) and, of these, about 90% become inactive at 98/08. As these clients have been selected by the agency through a scoring procedure, they are believed to be potentially good risks. Early evaluation of which segments of the portfolio are most likely to become inactive is an important information for the credit grantors. This involves prediction of intermediate variables, i.e. variables which are not known at the time the client is taken on file, that are of interest and expected to have an impact on the purely outcome variables. The conclusions are reliable as long as the selection strategy remains unchanged.

Graphical chain models seem natural condidates for modelling the context described. These are techniques to model situations where there is a so called *block-recursive* structure, that is a partial acyclic ordering among the variables, such that variables in the first group are merely explanatory, variables in the last group are merely outcome, and variables in the intermediate groups are response for the preceding ones and explanatory for following ones.

For each client a set of explanatory variables describing demographics and financial details is recorded. In this study we included a subset of these variables ("age", "number of children", "residential code" and "employment code") and put them in the first box of Figure 1. The variable in the second box concerns the information on the state of the account at the end of 1997. The last two boxes contain the average amount of loans and the state of the account at 98/08.

Although it is possible to build models for continuous and discrete variables, we decided to categorize the continuous variables. This has been done by first defining a thin categorization, and then by merging levels when the corresponding local odd ratios in the two-way table with the response variable (state of the account at 98/08) where close to one (i.e. in the interval 1 ± 0.01). A statistical procedure for merging levels of ordinal variables is in Cox and Wermuth (1998). In Table 1 the categorization used is presented. The variables included have a strong marginal association with the response variable.

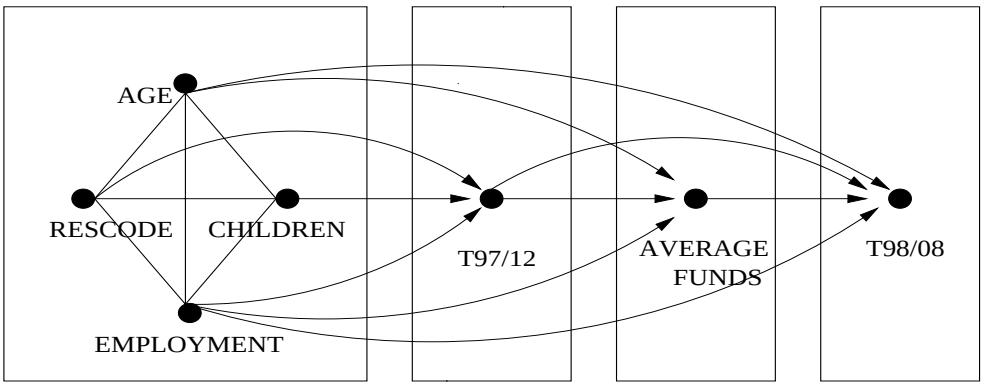

Figure 1: The chain graph for the behaviour of card holders.

The variables in the first box are used in the scoring system. The clients who entered the study have their score higher than a threshold. The selection strategy is therefore by conditioning on a response variable and the overall effect is to strengthen the interrelations among the explanatory variables.

One choice is therefore not to model their marginal distribution. However, we want to be able to draw conclusions also on subsets of population classified only on the basis of some of their demographic and financial characteristics. Therefore, the marginal model for them is of interest.

Conditional independence between pairs of ordinal variables has been tested with the Goodman and Kruskal test (Goodman and Kruskal, 1954, 1963) and between pairs of ordinal and nominal variables using Kruskal-Wallis test (see Agresti, 1984). Also, due to the sparsity of the table in the last two boxes, exact evaluation of the P-values for the likelihood ratios have been done. The ML estimate has been performed as described in Lauritzen (1996). The fitted chain graph is reported in Figure 1. The model is a result of a backward selection procedure. The P-value through the analyses has been set to 0.05.

For each intermediate variable, the graph shows the relevant variables. Inspection of the table obtained from the cross-classification of the clients according to those variables may lead to important conclusions. Here we summarize some of them.

Given the information on the residential status, the employment and the number of children, age has no direct impact on the behaviour at 97/12. The number of children has a strong impact on the probability of being active at 97/12 (in the expected direction) and indirectly influences the other two outcome variables. The impact of this variable on the state at 97/12 is, however, moderate for high income employees who are home owners.

For high level employees the probability that they are inactive at 98/08 is about 0.50 (to be compared with the marginal of 0.46). However, these people have a high probability (about 0.25, against the marginal of 0.18) of asking for loans of average amount in the last class. Therefore, they tend to use the card rarely and only for buying expensive goods. This situation becomes more marked for high income employees who are also home owners. In particular, if these clients are "dormant" at 97/12 they have a slightly higher probability than the average (about 0.92 against the marginal 0.90) of becoming "inactive" at 98/08. They may constitute an important population to target with marketing campaigns.

Self employed people tend to have an optimal profile. They show a higher probability than the population marginal in using the credit card (0.72 at 97/12 and 0.28 at 98/08, against, in order, 0.64 and 0.24) and if they are home owner, they also have a high probability (about 0.24) of asking for loans of average amount in the last class. Young people self employed who are dormant at 97/12, have a high probability of being active at 97/08 (0.10 against the marginal

of 0.06). They however concentrate on loans of small amount.

REFERENCES

Agresti, A. (1984). Analysis of Ordinal Categorical Data. New York: John Wiley & Sons.

Cox, D.R. and Wermuth, N. (1993). Linear Dependencies Represented by Chain Graphs. Statistical Science, 8, 3, 204-283.

Cox, D.R. and Wermuth, N. (1996). Multivariate Dependencies. Models, analysis, and interpretation. London: Chapman and Hall.

Cox, D.R. and Wermuth, N. (1998). On the Application of Conditional Independence to Ordinal Data. International Statistical Review, 66, 181-199.

Goodman, L.A. and Kruskal, W.H. (1954). Measures of Association for Cross Classifications. Journal of the American Statistical Associations, 49, 732-764.

Goodman, L.A. and Kruskal, W.H. (1963). Measures of Association for Cross Classifications III: Approximate Sampling Theory. Journal of the American Statistical Association, 58, 310-364.

Lauritzen, S.L. (1996). Graphical Models. Oxford: Oxford Science Publications.

Lewis, E.M. (1994). An introduction to Credit Scoring. San Raphael, California: The Athena Press.

Hand, D.J. and Henley, W .E. (1997) Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society, Series A, 160, 523-541.

Hand D.J., McConway K.J., and Stanghellini, E. (1996). Graphical models of applicants for credit. IMA Journal of Mathematics Applied in Business & Industry, 8, 143-155.

Sewart, P. and Whittaker, J. (1998). Graphical Models in credit scoring. IMA Journal of Mathematics Applied in Business & Industry, 9, 241-266.

Stanghellini E, Hand D.J. and McConway, K.J. (1999). A discrete variable chain graph for applicants for credit. Applied Statistics, 48, Part 2, 239-251.

Wermuth, N. (1992). On block-recursive linear regression equations (with discussion). Revista Brasileira de Probabilidade e Estatística, 6, 1-56.

RÉSUMÉ

Les crédits à la consommation représent un poids financier important dans nombreux pays industrialisés. Cette évolution incite les établissements accordant les crédits à développer des techniques permettant de discriminer entre les clients solvables et non-solvables. Aussi ces établissements ont-ils développé des produits nouveaux, non seulement pour élargir leur gamme de produits, mais surtout pour conserver des relations actives avec les clients solvables.
Le processus de discrimination entre le clients solvables et non-solvables implique néaumoins un nombre important de variables. L'objects de ce papier est de présenter une application des modèles graphiques dans tel contexte et d'analyser les advantages de cette approche.