ELSEVIER

International Journal of Forecasting 16 (2000) 149–172

international journal of forecasting

www.elsevier.com/locate/ijforecast

A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers

Lyn C. Thomas*

Department of Business Studies, University of Edinburgh, William Robertson Building, 50 George Square, Edinburgh EH8 9JY, UK

Abstract

Credit scoring and behavioural scoring are the techniques that help organisations decide whether or not to grant credit to consumers who apply to them. This article surveys the techniques used — both statistical and operational research based — to support these decisions. It also discusses the need to incorporate economic conditions into the scoring systems and the way the systems could change from estimating the probability of a consumer defaulting to estimating the profit a consumer will bring to the lending organisation — two of the major developments being attempted in the area. It points out how successful has been this under-researched area of forecasting financial risk. © 2000 International Institute of Forecasters. Published by Elsevier Science B.V. All rights reserved.

Keywords: Finance; Discriminant analysis; Classification; Economic forecasting; Profit scoring

1. Introduction

Forecasting financial risk has over the last thirty years become one of the major growth areas of statistics and probability modelling. When financial risk is mentioned one tends to think of portfolio management, pricing of options and other financial instruments (for example the ubiquitous Black–Scholes formula (Black & Scholes, 1973)), or bond pricing where Merton's paper (Merton, 1974) is seminal. Less well known but equally important are credit and behavioural scoring, which are the applications of financial risk forecasting to consumer lending. An adult in the UK or US is being credit scored or behaviour scored on average at least once a week as the annual reports of the credit bureaux imply. The fact that most people are not aware of being scored does not diminish from its importance. This area of financial risk has a limited literature with only a few surveys (Rosenberg & Gleit, 1994; Hand & Henley, 1997; Thomas, 1992, 1998) and a handful of books (Hand & Jacka, 1998; Thomas Crook & Edelman, 1992; Lewis, 1992; Mays, 1998). The aim of this survey is to give an overview of the objectives, techniques and difficulties of credit scoring as an application of forecasting. It also identifies two developments

*Tel.: +44-131-650-3798; fax: +44-131-668-3053.
E-mail address: l.thomas@ed.ac.uk (L.C. Thomas)

0169-2070/00/$ – see front matter
PII: S0169-2070(00)00034-0

in credit scoring where ideas from main-stream forecasting may help. Firstly there is a need to identify consumer risk forecasting techniques which incorporate economic conditions and so would automatically adjust for economic changes. Secondly, instead of seeking to minimise the percentage of consumers who default, companies are hoping they can identify the customers who are most profitable. Part of the catalyst for this development is the massive increase in information on consumer transactions which has happened in the last decade.

Credit scoring and behavioural scoring are the techniques that help organisations decide whether or not to grant credit to consumers who apply to them. There are two types of decisions that firms who lend to consumers have to make. Firstly should they grant credit to a new application. The tools that aid this decision are called credit scoring methods. The second type of decision is how to deal with existing customers. If an existing customer wants to increase his credit limit should the firm agree to that? What marketing if any should the firm aim at that customer? If the customer starts to fall behind in his repayments what actions should the firm take? Techniques that help with these decisions are called behavioural scoring

The information that is available in making a credit scoring decision includes both the applicant's application form details and the information held by a credit reference agency on the applicant. However there is also a mass of the information on previous applicants — their application form details and their subsequent performance. In many organisations such information is held on millions of previous customers. There is one problem with this information though. The firm will have the application form details on those customers it rejected for credit but no knowledge of how they would have performed. This gives a bias in the sample. This is a serious problem because if the firm says those it rejected previously would have been bad this decision will be perpetuated in any scoring system based on this data and such groups of potential customers can never have the opportunity to prove their worth. On the other hand there are usually sound reasons for rejecting such applicants and so it is likely that the rejects have a higher default rate than those who were previously accepted. Whether one can impute whether the rejected customers will be good or bad has been the subject of considerable debate. The idea of 'reject inference' has been suggested and used by many in the industry. Hsia (1978) describes the augmentation method while other approaches are suggested in Reichert, Cho and Wagner (1983) and Joanes (1993). Hand and Henley (1993) in a detailed study of the problem concluded that it cannot be overcome unless one can assume particular relationships between the distributions of the goods and the bads which hold for both the accepted and the rejected population. One way around it, is to accept everyone for a short period of time and to use that group as a sample. What firms do seems to depend as much on the culture of the organisation as on any statistical validation. Retailers and mail order firms tend to accept all applicants for a short period of time and use that group to build scorecards. Financial institutions on the other hand are swayed by the cost of default and feel there is no way they can accept everyone, even for a trial, and so use versions of reject inference.

In the next section we review the history of credit scoring. Then we examine the way credit scoring works and a general overview of the techniques that are useful in building credit scorecards. The fourth section gives a similar overview of behavioural scoring while the subsequent sections look at two proposed extensions of credit scoring which could give more robust and more focussed scorecards. The first extension tries to introduce dependence on economic conditions into credit scoring, while

the second is the change of objective from minimising default to maximising profit.

2. History of credit scoring

Credit scoring is essentially a way of recognising the different groups in a population when one cannot see the characteristic that separates the groups but only related ones. This idea of discriminating between groups in a population was introduced in statistics by Fisher (1936). He sought to differentiate between two varieties of iris by measurements of the physical size of the plants and to differentiate the origins of skulls using their physical measurements. David Durand (1941) was the first to recognise that one could use the same techniques to discriminate between good and bad loans. His was a research project for the US National Bureau of Economic Research and was not used for any predictive purpose. At the same time some of the finance houses and mail order firms were having difficulties with their credit management. Decisions on whether to give loans or send merchandise had been made judgementally by credit analysts for many years. However, these credit analysts were being drafted into military service and there was a severe shortage of people with this expertise. So the firms got the analysts to write down the rules of thumb they used to decide to whom to give loans (Johnson, 1992). These rules were then used by non-experts to help make credit decisions — one of the first examples of expert systems. It did not take long after the war ended for some folk to connect these two events and to see the benefit of statistically derived models in lending decisions. The first consultancy was formed in San Francisco by Bill Fair and Earl Isaac in the early 1950s and their clients at that time were mainly finance houses retailers and mail order firms

The arrival of credit cards in the late 1960s made the banks and other credit card issuers realise the usefulness of credit scoring. The number of people applying for credit cards each day made it impossible both in economic and manpower terms to do anything but automate the lending decision. When these organisations used credit scoring they found that it also was a much better predictor than any judgmental scheme and default rates would drop by 50% or more — see Myers and Forgy (1963) for an early report on such success or Churchill, Nevin and Watson (1977) for one from a decade later. The only opposition came from those like Capon (1982) who argued ‘that the brute force empiricism of credit scoring offends against the traditions of our society’. He felt that there should be more dependence on credit history and it should be possible to explain why certain characteristics are needed in a scoring system and others are not. The event that ensured the complete acceptance of credit scoring was the passing of the Equal Credit Opportunity Acts (ECOA, 1975, 1976) in the US in 1975 and 1976. These outlawed discriminating in the granting of credit unless the discrimination could be statistically justified. It is not often that lawmakers provide long term employment for any one but lawyers but this ensured that credit scoring analysis was to be a growth profession for the next 25 years. This has proved to be the case and still is the case. So the number of analysts in the UK has doubled even in the last four years.

In the 1980s the success of credit scoring in credit cards meant that banks started using scoring for their other products like personal loans, while in the last few years scoring has been used for home loans and small business loans. Also in the 1990s the growth in direct marketing has led to the use of scorecards to improve the response rate to advertising campaigns. In fact this was one of the earliest uses in the 1950s when Sears used scoring to decide to whom to send its catalogues (Lewis, 1992).

Advances in computing allowed other techniques to be tried to build scorecards. In the 1980s logistic regression and linear programming, the two main stalwarts of today's card builders, were introduced. More recently, artificial intelligence techniques like expert systems and neural networks have been piloted.

At present the emphasis is on changing the objectives from trying to minimise the chance a customer will default on one particular product to looking at how the firm can maximise the profit it can make from that customer. Moreover, the original idea of estimating the risk of defaulting has been augmented by scorecards which estimate response (how likely is a consumer to respond to a direct mailing of a new product), usage (how likely is a consumer to use a product), retention (how likely is a consumer to keep using the product after the introductory offer period is over), attrition (will the consumer change to another lender), and debt management (if the consumer starts to become delinquent on the loan how successful are various approaches to prevent default).

3. Overview of the methods used for credit scoring

So what are the methods used in credit granting? Originally it was a purely judgmental approach. Credit analysts read the application form and said yes or no. Their decisions tended to be based on the view that what mattered was the 3Cs or the 4Cs or the 5Cs:

- The character of the person — do you know the person or their family?
- The capital — how much is being asked for?
- The collateral — what is the applicant willing to put up from their own resources?
- The capacity — what is their repaying ability. How much free income do they have?
- The condition — what are the conditions in the market?

Credit scoring nowadays is based on statistical or operational research methods. The statistical tools include discriminant analysis which is essentially linear regression, a variant of this called logistic regression and classification trees, sometimes called recursive partitioning algorithms. The Operational Research techniques include variants of linear programming. Most scorecard builders use one of these techniques or a combination of the techniques. Credit scoring also lends itself to a number of different non-parametric statistical and AI modelling approaches. Ones that have been piloted in the last few years include the ubiquitous neural networks, expert systems, genetic algorithms and nearest neighbour methods. It is interesting that so many different approaches can be used on the same classification problem. Part of the reason is that credit scoring has always been based on a pragmatic approach to the credit granting problem. If it works use it! The object is to predict who will default not to give explanations for why they default or answer hypothesis on the relationship between default and other economic or social variables. That is what Capon (1982) considered to be one of the main objections to credit scoring in his critique of the subject.

So how are these various methods used? A sample of previous applicants is taken, which can vary from a few thousand to as high as hundreds of thousands, (not a problem in an industry where firms often have portfolios of tens of millions of customers). For each applicant in the sample, one needs their application form details and their credit history over a fixed period — say 12 or 18 or 24 months. One then decides whether that history is acceptable, i.e. are they bad customers or not, where a definition of a bad customer is commonly taken to be someone who has missed three consecutive months of payments. There will be a number of customers where it is not possible to determine whether they are good or bad because they have not been customers long enough or their history

is not clear. It is usual to remove this set of 'intermediates' from the sample.

One question is what is a suitable time horizon for the credit scoring forecast — the time between the application and the good/bad classification. The norm seems to be twelve to eighteen months. Analysis shows that the default rate as a function of the time the customer has been with the organisation builds up initially and it is only after twelve months or so (longer usually for loans) that it starts to stabilise. Thus any shorter a horizon is underestimating the bad rate and not reflecting in full the types of characteristics that predict default. A time horizon of more than two years leaves the system open to population drift in that the distribution of the characteristics of a population change over time, and so the population sampled may be significantly different from that the scoring system will be used on. One is trying to use what are essentially cross-sectional models, i.e. ones that connect two snapshots of an individual at different times, to produce models that are stable when examined longitudinally over time. The time horizon — the time between these two snapshots — needs to be chosen so that the results are stable over time.

Another open question is what proportion of goods and bads to have in the sample. Should it reflect the proportions in the population or should it have equal numbers of goods and bads. Henley (1995) discusses some of these points in his thesis.

Credit scoring then becomes a classification problem where the input characteristics are the answers to the application form questions and the results of a check with a credit reference bureau and the output is the division into 'goods' and 'bads'. One wants to divide the set of answers A into two subsets — $\boldsymbol{x} \in A_B$ the answers given by those who turned out bad, and $\boldsymbol{x} \in A_G$, the set of answers of those who turned out to be good. The rule for new applicants would then be — accept if their answers are in the set A_G; reject if their answers are in the set A_B. It is also necessary to have some consistency and continuity in these sets and so we accept that we will not be able to classify everyone in the sample correctly. Perfect classification would be impossible anyway since, sometimes, the same set of answers is given by a 'good' and a 'bad'. However we want a rule that misclassifies as few as possible and yet still satisfy some reasonable continuity requirement.

The simplest method for developing such a rule is to use a linear scoring function, which can be derived in three different ways — a Bayesian decision rule assuming normal distributions, discriminant analysis and linear regression. The first of these approaches assumes that:

- p_G is the proportion of applicants who are 'goods',
- p_B is the proportion of applicants who are bads,
- $p(\boldsymbol{x}|G)$ is the probability that a 'good' applicant will have answers $\boldsymbol{x}$,
- $p(\boldsymbol{x}|B)$ is the probability that a 'bad' applicant will have answers $\boldsymbol{x}$,
- $p(\boldsymbol{x})$ is the probability that an applicant will have answers $\boldsymbol{x}$,
- $q(G|\boldsymbol{x})(q(B|\boldsymbol{x}))$ is the probability that an applicant who has answers $\boldsymbol{x}$ will be 'good'('bad'), so
- $q(G|\boldsymbol{x}) = p(\boldsymbol{x}|G)\ p_G/p(\boldsymbol{x})$
- L is the loss of profit incurred by classifying a 'good' as a bad and rejecting them
- D is the debt incurred by classifying a 'bad' as a good and accepting them.

The expected loss is then:

$$L \sum_{\boldsymbol{x} \in A_B} p(\boldsymbol{x}|G)\ p_G + D \sum_{\boldsymbol{x} \in A_G} P(\boldsymbol{x}|B)\ p_B$$
$$= L \sum_{\boldsymbol{x} \in A_B} q(G|\boldsymbol{x})\ p(\boldsymbol{x}) + D \sum_{\boldsymbol{x} \in A_G} q(B|\boldsymbol{x})\ p(\boldsymbol{x}) \quad (1)$$

and this is maximised when the set of 'goods' is taken to be:

$$A_G = \{\boldsymbol{x}|Dp(\boldsymbol{x}|B)\ p_B \le Lp(\boldsymbol{x}|G)\ p_G\}$$
$$= \{\boldsymbol{x}|p_B/p_G \le (p(\boldsymbol{x}|G)L)/(p(\boldsymbol{x}|B)D)\}$$

If the distributions $p(\boldsymbol{x}|G)$, $p(\boldsymbol{x}|B)$ are multivariate normal with common covariance this reduces to the linear rule:

$$A_G = \{\boldsymbol{x}|w_1x_1 + w_2x_2 + \cdots\cdots w_mx_m > c\}$$

as outlined in several books on classification (Lachenbruch, 1975; Choi, 1986; Hand, 1981). If the covariances of the populations of the goods and the bads are different then the analysis leads to a quadratic discriminant function. However, in many classification situations (not necessarily credit scoring) (Titterington, 1992) the quadratic rule appears to be less robust than the linear one and the number of instances of its use in credit scoring is minimal (Martell & Fitts, 1981).

One could think of the above rule as giving a score $s(\boldsymbol{x})$ for each set of answers $\boldsymbol{x}$, i.e.

$$s(\boldsymbol{x}) = w_1x_1 + w_2x_2 + \cdots\cdots w_mx_m$$

If one could assume the discriminating power to differentiate between goods and bads was in the score $s(\boldsymbol{x})$ rather than in $\boldsymbol{x}$, then one has reduced the problem from one with m dimensions, represented by $p(\boldsymbol{x}|G)$, $p(x|B)$ to one with one dimension corresponding to the probabilities $p(s|G)$, $p(s|B)$. This is the power of a scoring system in that minimising the loss expression (1) reduces to finding the optimal cut-off for the score, namely:

$$\mathrm{Min}_c\{L\sum_{s<c} p(s|G)\ p_G + D\sum_{s\ge c} p(s|B)p_B\}$$

This simplification depends on the monotone behaviour of the inverse function $p(s|G)$ to ensure a unique optimal cut-off One can use various plots of score against probability of non-default to verify that the necessary conditions hold.

Returning to the general classification approaches to separating two groups (the goods and the bads in the credit scoring context), Fisher (1936) sought to find which linear combination of the variables best separates the two groups to be classified. He suggested that if we assume the two groups have a common sample variance then a sensible measure of separation is:

$M =$ (distance between sample means of two groups)/(sample variance of each group)$^{1/2}$

Assume that the sample means are $\boldsymbol{m_G}$ and $\boldsymbol{m_B}$ for the goods and the bads, respectively, and Σ is the common sample covariance matrix. If $Y = w_1X_1 + w_2X_2 + \cdots w_pX_p$, then the corresponding separating distance M would be:

$$M = \boldsymbol{w^T}\cdot(\boldsymbol{m_G} - \boldsymbol{m_B})/(\boldsymbol{w^T}\cdot\boldsymbol{\Sigma}\cdot\boldsymbol{w})^{1/2}$$

Differentiating this with respect to $\boldsymbol{w}$ and setting the derivative equal to 0 shows that this value M is minimised when $\boldsymbol{w}\propto(\Sigma^{-1}(\boldsymbol{m_G} - \boldsymbol{m_B}))$. The coefficients obtained are the same as those obtained in the Bayesian decision rule with multivariate normal distribution even though there has been no assumption of normality. It is just the best separator of the goods and the bads under this criterion no matter what the distribution. This follows since the distance measure M only involves the mean and variance of the distributions so gives the same results for all distributions with the same mean and variance.

The third way of arriving at the linear discriminant function is to define a variable Y equal to 1 if the applicant is good, 0 if the applicant is bad. The regression equation of the variable Y on the application form answers $\boldsymbol{X}$ gives a set of weightings on the predictive variables that agrees with that of the discriminant function, and this approach shows that the least squares approach of regression can be used to estimate the parameters. Myers and Forgy (1963) compared scorecards built using regression analysis and discriminant analysis, while

Orgler (1971) used regression analysis for recovering outstanding loans.

After the implementation of the Equal Credit Opportunities Acts, there were a number of papers critical of the discriminant analysis/regression approach (Eisenbeis, 1977, 1978). These criticised the fact the rule is only optimal for a small class of distributions (a point refuted by Hand, Oliver & Lunn (1996)). Others like Capon (1982) criticised the development and implementation of credit scoring systems in general because of the bias of the sample, its size, the fact that the system is sometimes overridden and the fact that there is no continuity in the score — so at a birthday someone could change their score by several points. These issues were aired again in the review by Rosenberg and Gleit (1994). Empiricism has shown though that these scoring systems are very robust in most actual lending situations, a point made by Reichert et al. (1983) and reinforced by experience (Johnson, 1992).

One feature of scorecard building whatever the technique used is that most of the application form questions do not give rise to numerical answers but to categorical ones, (do you own a phone; is your residential status that of owner, furnished renter, unfurnished renter or living with parents). There are several statistical methods for classifying when the data is categorical (Krzanowski, 1975; Vlachonikolis, 1986; Aggarawal, 1990). There are two ways credit scoring deals with these. One is to make each possible answer (attribute) to a question into a separate binary variable (Boyle et al., 1992; Crook, Hamilton & Thomas, 1992). Then the score for a consumer is the sum of the weights of the binary variables where the consumer's attributes have value 1. The problem with this is that it leads to a large number of variables from even a small number of questions. However, Showers and Chakrin (1981) developed a very simple scorecard for Bell Systems in this vein, in which the weights on all the answers were one — so one only had to add up the number of correct answers to get the score. Alternatively one can try and get one variable for each question by translating each answer into the odds of goods against bads giving that answer. Suppose 60% of the population are goods who own their phone, 20% are bads who own their phone, 10% are good with no phone, and 10% are bad with no phone. The odds of being good to being bad if you own a phone are $60/20 = 3{:}1$ or 3; the odds if you do not own a phone are $10/10 = 1{:}1$ or 1. So let the phone variable have value 3 if you own a phone, 1 if you do not. A slightly more sophisticated version is to take the log of this ratio which is called the weight of evidence, and is also used in deciding whether a particular variable should be in the scorecard or not. These approaches guarantee that within the variables, the different attributes have values with are in the correct order in terms of how risky that answer to the question is.

In fact these ways of dealing with categorical variables are also applied to the quantitative variables like age, income, years at present address. If one plots default risk with age (Fig. 1), one does not get a straight line (which would imply the risk is linear in age). One could all think of reasons why on reflection credit risk goes up in the mid-30s, but whatever it is this is a common phenomenon. Instead of trying to map such a curve as a straight line, one could either model it as a more complex curve or one could decide to group consumers into a number of categories and think of age as a categorical variable, which would allow the non-linearity to appear. The latter approach is the one commonly used in credit scoring mainly because one is already doing such groupings for the categorical variables. Here is where the art of credit scoring comes in — choosing sensible categories. This can be done using statistical techniques to split the variable so that the default risk is homogeneous within categories and is quite different in different categories. The classification tree tech-

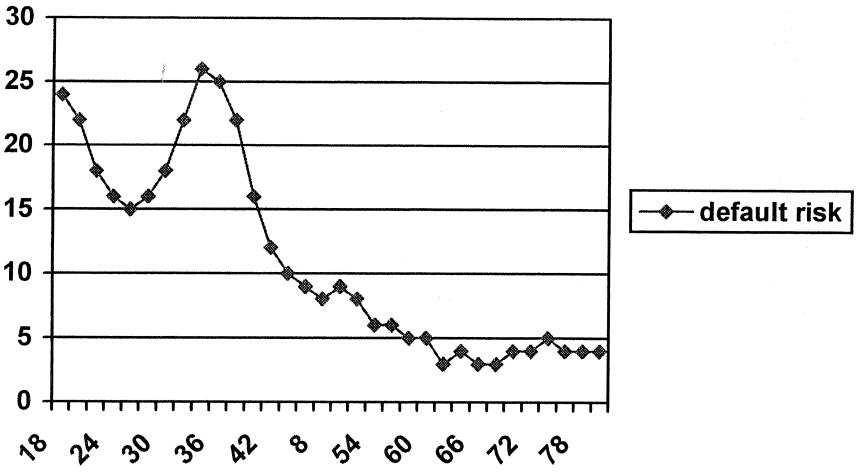

Fig. 1. Default risk against age.

niques which will be discussed later can be useful in doing this but is also important to consider life cycle changes when deciding on categories. Thus, in this age case one might choose 18–21, 21–28, 29–36, 37–59, 60+ — partly to reflect the change in statistics, partly because these are points where life cycle changes occur. Fig. 2 shows how the categories reflect the non-linear nature of risk with age.

The regression approach to linear discrimination says that p, the probability of default, is related to the application characteristics X_1, $X_2 \ldots X_m$ by:

$$p = w_0 + w_1X_1 + w_2X_2 + \cdots w_mX_m$$

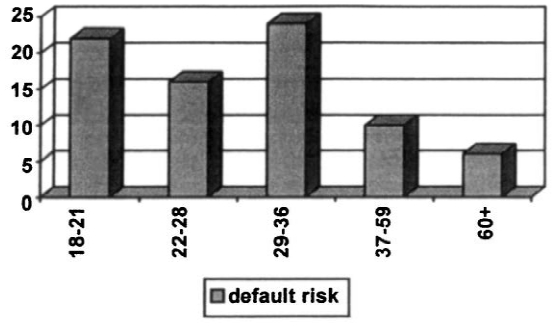

Fig. 2. Default risk against categorical age variables.

This has one obvious flaw. The right hand side of the above equation could take any value from $-\infty$ to $+\infty$ but the left hand side is a probability and so should only take values between 0 and 1. It would be better if the left hand side was a function of p which could take a wider range of values. One such function is the log of the probability odds. This leads to the logistic regression approach where one matches the log of the probability odds by a linear combination of the characteristic variables, i.e.

$$\log(p/(1-p)) = w_0 + w_1X_1 + w_2X_2 + \cdots + w_mX_m \quad (2)$$

Historically a difficulty with logistic regression was that one has to use maximum likelihood to estimate the weights w_i. This requires non-linear optimising techniques using iterative procedures to solve and is computationally more intensive than linear regression, but with the computing power available now this is not a problem Wiginton (1980) was one of the first to describe the results of using logistic regression in credit scoring and though he was not that impressed with its performance it has subsequently become the main approach to the classification

step in credit scoring. The actual classification results for linear and logistic regression are very similar and both are sensitive to correlations between the predictive variables and so one should try and ensure there are no strongly correlated variables left in the set on which the regression is calculated..

Eq. (2) implies that logistic regression can be thought of as linear regression where the dependent variable is some non-linear function of probability of being good. The score of the scorecard given by Eq. (2) is:

$$s(x) = w_0 + w_1X_1 + w_2X_2 + \cdots w_mX_m \qquad (3)$$

As explained earlier if the X_i are binary variables then the weights w_i are simply the score attached to that characteristic. If on the other hand, the X_i take other values usually related to the default risk of that attribute as outlined earlier, then the score for attribute i is w_ix_i.

In logistic regression, if one defines the scores as in Eq. (3) then the score relates to the probabilities by:

$$\begin{aligned} s(\boldsymbol{x}) &= \ln(p(G|\boldsymbol{x})\ p(B|\boldsymbol{x}) \\ &= \ln\{(p_G/p_B)(p(\boldsymbol{x}|G)/p(\boldsymbol{x}|B))\} \end{aligned}$$

This shows that the logistic approach is directly estimating the information odds $(p(\boldsymbol{x}|G)/p(\boldsymbol{x}|B))$ rather than making assumptions about $p(G|\boldsymbol{x})$. In fact, Fung, Lucas, Oliver and Shikaloff (1997) initially assume independence of the information odds for different characteristics, i.e.

$$\begin{aligned} s(\boldsymbol{x}) &\approx (p(\boldsymbol{x}|G)/p(\boldsymbol{x}|B)) \\ &= (p(x_1|G)/p(x_1|B))(p(x_2|G)/p(x_2|B))\cdots \\ &\quad \cdots(p(x_m|G)/p(x_m|B)) \end{aligned}$$

and then use a recursive procedure for improving the estimates

This discussion on whether $p(\boldsymbol{x}|G)$ or $p(G|\boldsymbol{x})$ is the basic quantity being estimated highlights the role of the population odds (p_G/p_B) in transforming one to the other. These could be estimated for the population as a whole but this is rarely done in practice. Instead this estimation is hidden away in the choice of a suitable cut-off. Normally this choice of cut-off and hence population odds, is done using the hold-out samples.

Another non-linear regression is probit analysis suggested by Grablowsky and Talley (1981). In probit analysis if $N(x)$ is the cumulative normal distribution function so:

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\mathrm{d}y$$

then the aim is to estimate $N^{-1}(p_i)$ as a linear function of the characteristics of the applicant, so:

$$\begin{aligned} N^{-1}(p_i) &= w_0 + w_1X_1 + w_2X_2 + \cdots\cdots \\ &\quad + w_mX_m \end{aligned}$$

This is equivalent to saying that someone is good if their score is above a certain level, but this level varies from individual to individual and has a normal distribution.

Linear programming used as a classification approach also ends up with a linear scorecard. Suppose one has a sample of n_G goods and n_B bads and a set of m predictive variables from the application form answers so borrower i has predictive variable values $(x_{i1}, x_{i2}, \cdots, x_{im})$. One seeks to develop a linear scorecard where all the goods will have a value above a cut-off score c and all the bads have a score below the cut-off score. This cannot happen in all cases so we introduce variables a_i which allow for the possible errors — all of which are positive or zero. If we seek to find the weights $(w_1, w_2, \cdots w_m)$ that minimise the sum of the absolute values of these errors we end up with the following linear programme:

$$\text{Minimise } a_1 + a_2 + \cdots\cdots + a_{n_G+n_B}$$

subject to

$$w_1x_{i1} + w_2x_{i2} + \cdots\cdots + w_mx_{im} \geq c - a_i \quad 1 \leq i \leq nG$$

$$w_1x_{i1} + w_2x_{i2} + \cdots\cdots + w_mx_{im} \leq c + a_i \quad n_G + 1 \leq i \leq n_G + n_B$$

$$a_1 \geq 0 \quad 1 \leq i \leq n_G + n_B$$

Other approaches allow one to minimise the maximum error — change a_i to a in each constraint. Mangasarian (1965) was the first to recognise that linear programming could be used for discrimination, but it was the papers by Freed and Glover (1981a,b) that sparked off the interest. The subsequent substantial literature on the subject is reviewed by Joachimsthaler and Stam (1990). Although Nath, Jackson and Jones (1992) found that statistical methods were superior to linear programming ones, others have found that LP classifies as well as statistical approaches, including Hardy and Adrian (1985) who looked at credit scoring applications. Latterly there has been more work on using integer programming to solve problems with different ways of describing misclassification error (Glen, 1997), or using hybrid schemes involving both linear programming and statistical methods (Ziari, Leatham & Ellinger, 1997). One of the more famous of the integer programming applications in credit scoring is the AT&T scorecard developed by Kolesar and Showers (1985) mentioned earlier where there was a prerequirement that the scorecard be of a very simple form — just count the number of 'yes' answers in the application form.

Classification trees or recursive partitioning algorithms (RPA) and expert systems do not end up with a scorecard which gives weights to each answer and then adds these weights. Instead they classify the consumers into groups, each group being homogeneous in its default risk and as different from the default risks of other groups as is possible. Classification trees have been developed in statistics (Breiman, Friedman, Olshen & Stone, 1984), and in artificial intelligence (Safavian & Landgrebe, 1991), and in machine learning (Quinlan, 1993). In classification trees one splits the set of application form answers into two subsets. Fixing on the answer to one question, one chooses the split of possible answers to the question into two subsets where the difference in average default risk between the two subsets is as large as possible. Other criteria are less myopic and look ahead to the situation after several more levels of splits. Having found the best split for a specific question, the process is repeated for all the questions. One chooses the question and the split of its answers that maximises the difference in default risk between the two subsets created. One then repeats this process on each of the two subsets in turn. One keeps on splitting subsets of the consumers until either one ends up with groups which are so small that it is not statistically sensible to split anymore or that the best split produces two new subgroups which are not statistically significantly different. When one has stopped splitting the tree, one classifies each remaining group as good or bad depending on whether the majority in that group are good or bad. Fig. 3 gives an example of such a tree.

One has to prune back the tree obtained to get a scheme that is more robust in classifying other samples even if it is not so accurate on the one it was developed on. There are alternative approaches to making the classification trees robust like averaging over several large trees but pruning is by far the most common approach. Makowski (1985) was one of the first to advertise the use of classification trees in credit scoring, whereas Coffman (1986) compared trees with discriminant analysis and suggested that the former is better when there is interaction between the variables and the latter when there is intercorrelations. Mehta (1968), Carter and Catlett (1987) and Boyle et al. (1992) discuss the results of using classification trees in credit scoring. More recently there have been investigations of oblique trees where there is not

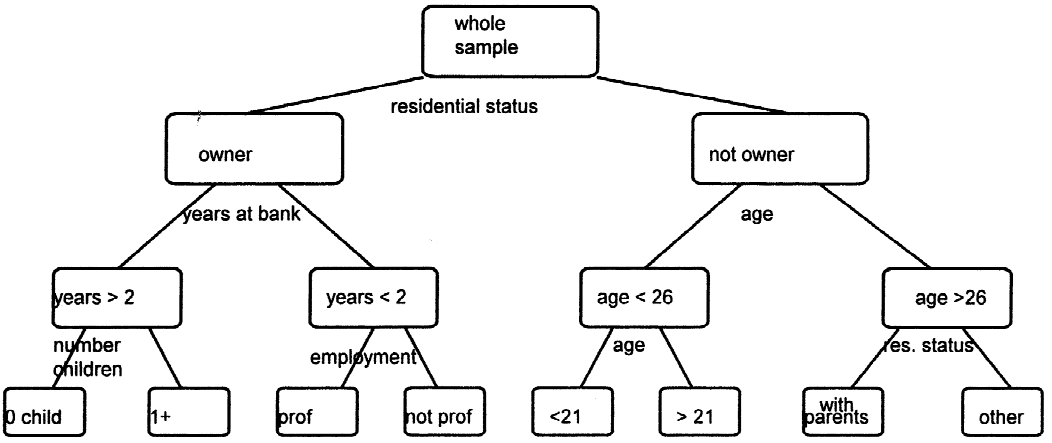

Fig. 3. Classification tree for credit scoring.

a clean division between the two branches at a node but there may be some common elements

There are also four other approaches to credit scoring, which have been piloted in the last decade without becoming fully established. They all lead to classification systems which are not linear scorecards. Neural networks, which can be considered as a form of non-linear regression (Cheng & Titterington, 1994) have proved an ubiquitous approach to many problems and are highly suited to credit scoring applications. Most credit applications of neural networks have been to the scoring of corporations, where there is less data available, than in scoring of consumers (Altman, Marco & Varetto, 1994; Tam & Kiang, 1992). Desai, Crook and Overstreet (1996) and Desai, Conway, Crook and Overstreet (1997) compared neural networks with regression and genetic algorithm approaches for consumer credit scoring in the US credit union environment where again there are fewer customers than in normal credit card situations. In other types of applications, hybrid approaches using neural nets with other classifying techniques have been used. For example, Mangasarian (1993) used linear programming to train the neural nets. Ignizio and Soltys (1996) produce masking functions which cover the region of one type of credit risk by an amalgam of neural networks and linear programming.

Neural networks and classification trees are sometimes called expert systems as they are automated procedures that have learning abilities. Expert system also describes systems where the human experts' learning has been incorporated into a set of rules, some of which have been developed using an inference engine from data presented to the system. Again most of the credit applications of this technique have been in commercial loan situations (Zocco, 1985; Leonard, 1993a) — or fraud detection (Leonard, 1993b). Tessmer (1997) described how inductive learning can be used in the consumer credit granting problem. Davis, Edelman and Gammerman (1992) looked at how a Bayesian expert system could work on classifying credit card applicants and compared the results with a neural net approach. However, the real successes of expert systems in the credit environment has been in detecting fraud at the transaction stage (Davis, 1987).

Genetic algorithms are one of a number of general optimization schemes based on biological analogies. In the credit scoring context one

has a number of scorecards which mutate and blend together according to their fitness at classification. Fogarty and Ireson (1993) and Albright (1994) were one of the first to describe this approach. Desai et al. (1997) compared it with neural networks in the credit union environment, while Yobas, Crook and Ross (1997) did a comparison of these two and classification trees using credit card data.

Nearest neighbours, a non-parametric statistical approach, has an obvious appeal in the credit scoring environment. One chooses a metric on the space of application data to measure how far apart applicants are. All the applicants in a sample of past data then have a position in this metric space. A new applicant is then classified according to which group — good or bad — is in the majority among the nearest neighbours to the new applicant. The analysis of Henley and Hand (1996) suggests that the classification is fairly robust to the choice of how many neighbours should be considered, and the system has the advantage that new data points can be added and so the system be updated with no change to the underlying coding. Chatterjee and Barcun (1970) were one of the first to suggest this approach to credit scoring.

So which method is best? Each consultancy says its approach is best, while the comparisons by academics are often limited as some of the most significant data like the credit bureau reports are too sensitive or too expensive to be passed to them by the users. Thus their results are only of an indicative nature but generally there is only a small change between the classification errors of different methods. Table 1 shows the results of five comparisons. The numbers should be compared across rows but not between rows because they involve different measures of good on different populations. They each give % correctly classified by the different methods when the same proportion of the population is accepted by each method. In the Henley and the Srinivasan comparisons RPA is just the winner; in the Boyle and Yobas papers it is linear regression and in the Desai it is logistic regression, but the results are not significant.

The reason for this may be the flat maximum effect first identified by Lovie and Lovie (1986) 20 years ago. This says that significant changes in the weights around the optimal scorecard have relatively little effect on its performance. It would explain the relative similarity in results of very different scorecard building approaches. This flat maximum effect prompted experts to wonder if scorecards are robust to differences in the population being scored. Could one have a generic scorecard where one builds a scorecard on one group of consumers and uses it to score a geographically or socio-economically different group of consumers. One such experiment was to try and build an European scorecard (Platts & Howe, 1997) which can be used in all the countries of Europe. Another was to build a generic scorecard for credit unions in the US

Table 1
Comparison of classification accuracy for different scoring approaches

Authors	Linear reg	Logistic reg	RPA	LP	Neural nets	GA
Henley (1995)	43.4	43.3	43.8	–	–	–
Boyle et al. (1992)	77.5	–	75	74.7	–	–
Srinivasan and Kim (1987a,b)	87.5	89.3	93.2	86.1	–	–
Yobas et al. (1997)	68.4	–	62.3	–	62.0	64.5
Desai et al. (1997)	66.5	67.3	67.3	–	6.4	–

(Overstreet, Bradley & Kemp, 1992) who usually only have a small numbers of clients. In both cases the results are better than not using a scorecard but do not seem to be competitive with tailored scorecards. This suggests the differences in the populations in different countries and in the case of credit unions in different occupational groups do affect the scorecard performance significantly.

So if different methods give about the same level of classification accuracy which one should be used? This is determined by the other features the methods bring to scorecard building. The regression approaches allow one to perform statistical tests to identify how important are each of the application form questions to the accuracy of classification and whether two different questions are essentially asking the same thing and getting equivalent responses. Thus these approaches allow one to drop unimportant questions, which makes the scorecards more robust — they are lean and mean. They also indicate which are the important questions for classification purposes and hence help in deciding what questions to ask in new scorecards.

Linear programming has the advantage that it allows the scorecard designer to ensure that scores have properties that the lending organisations may want. For example, the lender might want to target young people and so want the scores for being under 25 to be greater than that for the over 60s. Finding the best scorecard with this property is quite hard for the statistical approaches but very easy if one uses linear programming. Similarly LP can deal with a lot more variables than the regression approaches can and so copes easily when categorical questions are translated into many binary variables. Classification trees and neural nets are good at automatically finding the non-linear relationships between the variables which cause errors in linear scorecards if they are not recognised.

4. Overview of behavioural scoring

Behavioural scoring systems allow lenders to make better decisions in managing existing clients by forecasting their future performance. The decisions to be made include what credit limit to assign, whether to market new products to these particular clients, and if the account turns bad how to manage the recovery of the debt. The extra information in behavioural scoring systems compared with credit scoring systems is the repayment and ordering history of this customer. Behavioural scoring models split into two approaches — those which seek to use the credit scoring methods but with these extra variable added, and those which build probability models of customer behaviour. The latter also split into two classes depending on whether the information to estimate the parameters is obtained from the sample of previous customers or is obtained by Bayesian methods which update the firm's belief in the light of the customer's own behaviour. In both cases the models are essentially Markov chains in which the customer jumps from state to state depending on his behaviour.

In the credit scoring approaches to behavioural scoring one uses the credit scoring variables and includes others which describe the behaviour. These are got from the sample histories by picking some point of time as the observation point. The time preceding this — say the previous 12 months — is the performance period and variables are added which describe what happened then — average balance, number of payments missed. etc. A time some 18 months or so after the observation point is taken as the performance point and the customer's behaviour by then is assessed as good or bad in the usual way. Hopper and Lewis (1992) give a careful account of how behavioural scoring systems are used in practice and also how new systems can be introduced.

They advocate the Champion vs. Challenger approach where new systems are run on a subset of the customers and their performance compared with the existing system. This makes the point yet again that it takes time to recognise whether a scoring system is discriminating well.

The choice of time horizon is probably even more critical for behavioural scoring systems than credit scoring systems. Behavioural scoring is trying to develop a longitudinal forecasting system by using cross-sectional data, i.e. the state of the clients at the end of performance period and at the end of the outcome period. Thus the time between these periods will be crucial in developing robust systems. Experimentation (and data limitations) usually suggest a 12 or 18-month period. Some practitioners use a shorter period, say 6 months, and then build a second scoring system to estimate which sort of behaviour at six months will lead to the client eventually defaulting and define this 6 month behaviour as 'bad' in the main scorecard. One can use older data for the second scorecard while using almost current data for the main scorecard.

The probability models classify the different states the consumer can be in using variables from the application form and variables describing current and recent behaviour, for example — balance outstanding, number of periods since a payment was made, average balance. The following example takes this approach to a revolving account where a customer is both paying for previous orders and ordering new items.

Let the states, which describe the customers account be given by $u = (b,n,i)$, where b is the balance outstanding, n is the number of periods since the last payment and i is any other relevant information. Suppose the action is which credit limit, L, to set and we assume the performance of the account may be affected by the credit limit set. It is necessary to estimate $p^L(u,u')$ and $r^L(u)$, where $p^L(u,u')$ are the probability of the account moving form state u to u' under a credit limit L in the next period and $r^L(u)$ is the likely reward obtained in that period. These can be obtained by estimating $t^L(u,a)$, the probability that an account in state u with credit limit L repays a next period; $q^L(u,o)$, the probability that an account in state u with credit limit L orders o next period; $w^L(u,i')$, the probability that an account in state u with credit limit L changes its information state to i' and defining transition probabilities by:

$$p^L(b,n,i;\, b+o-a,0,i') = t^L(u,a)\, q^L(u,o)\, w^L(u,i'),$$
provided $b+o-a \le L$, and $a > 0$

$$p^L(b,n,i;\, b-a,0,i') = t^L(u,a)\, w^L(u,i')(q^L(u,0) + \sum_{o>L-b+a} q^L(u,o)),$$
where $a > 0$

$$p^L(b,n,i;\, b+o,n+1,i') = t^L(u,0)\, q^L(u,o)\, w^L(u,i'),$$
provided $b+o \le L$

$$p^L(b,n,i;\, b,n+1,i') = t^L(u,0)\, w^L(u,i')(q^L(u,0) + \sum_{o>L-b+a} q^L(u,o))$$

If f is the fraction of a purchase that is profit for the company and the company has a policy of writing off bad debt after N periods of non-payment that the reward function would be

$$r^L(b,n,i) = f \sum o\, q^L(u,o) - b t^L(u,0)\, \delta(n-(N-1))$$

One can then use dynamic programming to find $V_n(u)$ the expected profit over n periods given the account is in state u and the optimal credit limit policy by solving the optimality equation:

$$V_n(u) = \max_L \{r^L(u) + \sum_{u'} p^L(u,u')\, V_{n-1}(u')\}$$

The first published account of this type of model was by Cyert, Davidson and Thompson (1962), where the units were dollars not accounts and the state was how overdue the account was. Their approach had some difficulties with accounting conventions — an account with £10 three months overdue and £10 one month overdue would become four months overdue if only £10 is paid in the next month. This pioneering paper was followed by several which modified the basic model. van Kuelen et al. (1981) suggested a modification of the approach that overcame the difficulty with defining partial payments of overdue accounts while Corcoran (1978) pointed out that the system would be even more stable if different transition matrices were used for accounts of different characteristics such as size of the accounts, i.e. $p(b,n,i;\ b',n',i')$ really does depend on the other factors i. Frydman et al. (1985) split the population into 'movers' and 'stayers', where the latter are more likely to stay in their current state (usually the paid up to date state). The question on how many segments of the population should have different scoring systems is important in credit scoring as well as behavioural scoring. Banasik et al. (1996) point out that segmentation does not always give an improved scorecard in practice, if the segments are not distinctive enough.

An alternative Bayesian based probability model was pioneered by Bierman and Hausman (1970). In this the probability of paying was not given from a sample of previous customers but was taken to be a Bernoulli random variable whose parameter satisfied a Beta distribution. The parameters of the Beta distribution were updated by the payment performance of the individual customer, so if initially they were (r,n) than after n' payments periods in which the customer paid r' times they became $(r+r', n+n')$. The authors assumed that once credit had been refused no more credit was granted, unlike the model described earlier in this section. Dirickx and Wakeman (1976) relaxed this assumption, while Srinivasan and Kim (1987a,b) allowed the simple extension of payments and orders being possible in the same period. Thomas (1994) extended the model by allowing not only the probability of repayment but also the maximum affordable repayment amount to be random variables which are updated in a Bayesian fashion according to the amount of repayments made.

5. Incorporating economic conditions into credit and behavioural scoring

Credit scoring is now used in almost all forms of consumer lending — credit cards, personal loans, car finance, insurance policies, utility payments. For the last decade, mortgage scoring has been successfully applied in the US. The connection between credit scoring and response scoring (to see who is likely to respond to direct marketing campaigns) is being blurred as organisations seek to market to people they know they will accept for lending products. This blurring is an area of concern for the data protection lobby. Banks have begun to recognise that lending £10,000 on a credit card to the owner of a one-man business and lending £10,000 to his firm is a similar sort of decision. So scorecards are being developed for lending to small businesses (Edelman, 1997). In the US credit scoring is used to estimate the risk of a portfolio of consumer debt which one financial organisation may want to sell off to another. Moreover, with a good risk estimating instrument it is possible to introduce risk based pricing, though it is surprising how slowly risk based pricing is taking to develop in lending products given its long history in terms of no claims bonuses in car insurance.

In all these applications one important advance would be to incorporate economic conditions into scoring methodologies. There can

be a several year time lag between the transaction data collected and its use in a scorecard. A scorecard in use in 1999 was likely to be built in 1998. In order to have enough history to decide if the customers in a sample were good or bad, a sample of customers who joined in 1995 or 1996 would be needed. This caused a real problem in the recession at the beginning of the 1990s when the architects and accountants who had wonderful credit histories in 1988 and 1989 were the ones who were defaulting on their loans in 1991. Thus scorecard have to be constantly redeveloped — around every 18 months to 2 years in order to overcome this drift in the population. To give an example of the changes that occur even in one year, Crook, Hamilton and Thomas (1992) built two scorecards on a sample of consumers for the same lending product — one built on customers joining in 1988 using their credit history in 1989 — relatively good economic conditions — and one on customers joining in1989 using their history in 1990 — when conditions were worse. Both scorecards were used to score both sets of consumers. The results are given in Table 2. If one adjusts the cut-off to accept the same % in each year (a cut-off at the same marginal odds of goods to bads would take far less on the 1990 scorecard) — one finds that 25% of the group who would be rejected in one year would be accepted in the other and vice versa.

This is not to say that economic conditions are the only cause of changes in risk behaviour. Zandi (1998) looked at the US experience in the 90s where there was a massive increase in consumer defaults and bankruptcy in 1995–1997 even though the economy kept improving. He put this down to the lowering of credit standards in the previous few years as financial organisations competed for credit card customers and home loan borrowers by dropping their cut-off levels. His regression analysis of personal bankruptcy on economic variables including unemployment claims did show though that economic conditions do have a major impact on default.

Table 2
Swap sets between scorecards built on data 1 year apart

Scorecard built on 1989 (good year) data	Scorecard built on 1990 (bad year) data	
	Accepted	Rejected
Accepted	79.9%	3.7%
Rejected	3.7%	12.8%

So how can you build in economic conditions to the scorecard? One way would be to build scorecards for different economic conditions so a customer would have a score for good times and one for bad times. However, this puts the onus back on the credit manager to decide what is the future and what is a suitable decision rule. Also, the data could be very old if you want to build a score for all the stages of an economic cycle with the problem that there might be socio-demographic changes in the population in this time as well as changes in the economic conditions. In a sense one is applying log linear models to economic and application variables and their interactions and one of the difficulties with log linear models with interactions is how quickly one runs out of data. Zandi (1998) described a simpler model where one adds on to the normal credit score, a score based on leading economic indicators for that customer, which is based on the geographic area and employment type of the customer.

A way of trying to build tighter models to connect economic conditions, application variables, and consumer behaviour is to use the graphical methods and Bayesian learning networks. Sewart and Whittaker (1998) and Hand, McConway and Stanghellini (1997) pointed out how useful these techniques are for examining the relationships between predictive and outcome variables. Fung et al. (1997) showed how using the ideas of cliques and Markov blankets

fits very well into the credit scoring context since the analysis leads to a scoring system where the good–bad odds score is additive over the cliques. This type of analysis could be expanded to include economic variables and so lead to relationships between the outcome variables and the predictive and economic variables.

In behavioural scoring one way of dealing with economic conditions is not to try and introduce these into the score itself but rather into the dynamics of the way the score will change. Thus one could say that the transition probabilities of the Markov chain which represent the behavioural score are in fact dependent on the economic conditions, which in turn could be modelled as a Markov chain. These hidden Markov chain models have proved one way of modelling economic conditions in the related problem of pricing corporate bonds using their credit ratings (Thomas, Allen & Morkel-Kingsbury, 1998).

6. Profit scoring

The other major change in the last few years is that credit lenders wish to change from minimising the risk of a consumer defaulting to maximising the profit a consumer brings them. Initially one may feel that all is required is a change in the definition of 'good' in the previous techniques and some organisations have gone along this path. However whereas default rates are affected by acceptance decisions, credit limit decisions and default recovery decisions, profits are affected by many more decisions including marketing, service levels and operation decisions as well as pricing decisions. Thus moving to profit scoring implies that these techniques should help a whole new range of decisions — in fact almost all the decisions a retailer or retail bank may be involved in.

There are a number of implementation problems encountered in making this change to profit scoring, which is why it is taking organisations so long to move to fully blown profit scoring systems. Firstly, there are data warehousing problems in ensuring the accounts include all the elements which make up the profit. Even in credit card organisations this has proved difficult in that the credit card company gets a certain percentage of each purchase made on the card paid back to it by the retailer — the merchant service charge. This charge varies considerably between the types of purchases and the companies have had to revamp their systems so that this information can be readily accessed. Similarly the retail part of an organisation often writes off all or a fixed percentage of any bad debt a customer incurs and never checks how much of it is actually recovered subsequently by the debt-recovery department. These examples suggest profit scoring requires a fully integrated information system in the organisation. One needs the information on all the customers' transactions (and maybe a whole family's transactions) and accounts collated together in order to calculate the customers' profitability to the firm. Hence, the push to data warehousing by companies so that all this information is kept and is easily accessible. This could lead to legal problems as the use of personal information for reasons other than those for which it was originally collected is frowned upon by legislators in many countries.

The advent of data mining techniques (see Jost, 1998 for their uses in credit scoring), mean that the technical problems of analysing such vast amounts of data are being addressed. However, there are still major problems in developing models for profit scoring. What is a reasonable time horizon to consider profit over which prevents a strategy of alienating customers by high prices now and forgets about the future? Also, profit is a function of economic conditions as well as the individual consumer's characteristics. So it is even more important to

include economic variables into profit scoring than it was in credit scoring. Profit is dependent on how long a customer stays with a lender and so one wants to know how long customers stay and whether they default or just move their custom elsewhere. So one needs to estimate attrition rates as part of profit scoring

Lastly, there are two difficulties that affect which methodology to choose. Should one look at the profit on each product in isolation or look at the total profit over all possible products. The former means one could decide not to offer a customer a new credit card because he does not use it enough. This refusal may offend a customer so much that his profitable home loan is moved to another lender. Going for total profit on the other hand ignores the fact that the decision on which product a customer takes is the customer's decision. He can cherry pick and may refuse the product where the firm felt it would make the most profit from him. Secondly, there is the problem of censored data. In a sample of past transactions, the total profit for current customers will not be known, but only the profit up to the date that the sample history finished.

So what approaches are being tried and what approaches might work. We classify them into four groups. One approach is to build on the existing scorecards which estimate default, usage, acceptance and attrition and try to define the profit for groups of the population segmented according to their scores under these different measures. Oliver (1993) was one of the first to suggest this and looked at what decision rules should be used if one has a 'transaction profit' score and a default score. Fishelson-Holstine (1998) described a case study where one tried to segment according to two types of profit. A bank runs a private label credit card for a retailer. The retailer wants to increase profits by sending cards to people who will use it to buy more in their stores, while the bank wants the credit card operations of the customer to be profitable. By using a demographically based segmentation tool, details of the retailers sales and the credit card transaction database, groups were identified who were profitable for both. Li and Hand (1997) suggested an intermediate approach where instead of trying to estimate the final profit or default criterion directly, one should try to estimate intermediate variables like balance outstanding, purchases, etc and use these to estimate the final outcome. Simulation studies suggested this approach was not necessarily superior to estimating the final outcome directly. This approach could benefit from using the Bayesian learning/graphical network tools (Hand et al., 1997; Sewart & Whittaker, 1998) described earlier to identify how default, usage, acceptance and attrition should be combined to reflect profit.

A second approach is to mimic the regression approach of credit scoring by trying to describe profit as a linear function of the categorical application form variables. Almost all the data will be censored in that the total profit is not known but there is a body of literature on regression with censored data (Buckley & James, 1979). Research in this area is continuing (Lai & Ying, 1994) but the type of censoring that occurs in credit scoring has not yet been dealt with satisfactorily.

The third approach is to build on the Markov chain approaches to behavioural scoring to develop more precise stochastic models of customer behaviour. Cyert et al.'s (1962) original model could be used to model profit in a one product case and these approaches have proved very successful in estimating debt provisioning for portfolios of customers with the same product. If one extends the ideas to the total profit over several products, the problem becomes one of data availability and computational power. One runs into the 'curse of dimensionality' that arises when one seeks to use Markov chains to model complex real situations. However, there are a number of techniques that are proving

very successful in other application areas which overcome this problem (Thomas, 1994).

Fourthly, one could recognise that another area where there has been successful statistical modelling with lots of censored data is survival analysis which estimates the reliability of machines and people. One could try to use the techniques of this area — proportional hazards models, and accelerated life models — to estimate the long term profit from a customer given only the experience in the first few months or years. Narain (1992) was the first to suggest that one could use this analysis on credit scoring data, while the paper by Banasik et al. (1999) showed that one could also use the idea of competing risks from reliability to get good estimates of when borrowers will default and when they will pay off early, thus incorporating default and attrition in the same analysis. The accelerated life approach is also a useful way of thinking about how economic affects can be introduced into profit models. By taking the characteristic variables in proportional hazards and accelerated life to describe the economic conditions as well as the characteristics of the borrower one can build a model that allows for the 'speeding up' in default rates that occurs in poor economic conditions. This technique was used in estimating the default rates of corporate bonds by Lando (1994) in his PhD thesis.

Profit scoring systems seem more difficult to obtain than might have been first thought, but the prize for success would be enormous. It would provide a decision support aid that has the same focus throughout the different decision making areas of the organisation. It would provide an excellent way of taking advantage of all the new data on consumer behaviour that has become available in the last few years with electronic point of sales equipment and loyalty cards. With the investment in data warehousing and data mining packages, organisations now have the capability to classify and segment all this information. Profit scoring would provide the objectives and models to use this information.

7. Conclusions

This review seeks to give an overview of the techniques that are used and being developed to forecast the financial risk involved in lending to consumers. Previous surveys have concentrated only on statistical approaches or restricted themselves to the initial credit granting decision while we seek to cover both credit and behavioural scoring. We have also sought to give a fairly comprehensive biography of the literature of the topics we cover.

Credit and behavioural scoring have become establishes as major tools in forecasting financial risk in consumer lending and in helping organisation cope with the risk of default in consumer lending. Once an organisation takes up statistically and Operational Research based credit scoring, it hardly ever returns to judgmental based ones (Lewis, 1992). In practice, the fears of Capon (1982) and the difficulties alluded to in Rosenberg and Gleit (1994) have been allayed. As scoring usage expands to newer area — mortgage scoring for example — there may be reasons why it should be combined with judgmental systems or ones based on 'loan to value' of the secured item, which traditionally has proved successful. The organisation needs to identify what risk it wishes to protect against and whether scoring is the appropriate technique of quantifying that risk.

There are social issues in using scoring as a forecasting tool. It is illegal to use some characteristics — race, sex, religion — but that does not prevent some authors (Crook, 2000) suggesting that there are surrogate variables which mean scoring systems do discriminate in these areas. Other authors (Chandler & Ewert, 1976) argue the relationship of these banned charac-

teristics with other allowed characteristics forces allows the very discrimination which one is seeking to avoid. We have left discussion of these questions as well as the methods for calibrating the effectiveness of a scorecard out of this review, since both are major topics in their own right.

In this review we have speculated on two areas in which there is a need for major developments in the models and techniques available. Progress in incorporating economic effects would mean scorecards would be more robust to changes in the economic environment and so could be used for longer time periods before having to be rebuilt. Profit scoring would allow organisations to have a tool that is more aligned to their overall objective than the present tools which estimate the risk of consumers defaulting. However, if these developments are successful there may well be major impacts on the credit industry and on consumers. For the industry, those with the best models of consumer behaviour will make the best profit — so there will be strategic advantages in having models which best analyse the wealth of data coming through. Firms, who are confident in their models, will start cherry picking and going for the most profitable customers. The subsequent price changes will lead to a levelling of the profits, but it will also lead to a standardisation between financial and retail organisations about the types of consumers they want. Thus some people will be able to borrow from all and will be the target of most organisations, but there may be an underclass of consumers who cannot borrow — certainly not in the plastic card market — and who are not targeted for any marketing. With lending and retailing becoming more automated, these consumers will face growing disadvantages and this may lead to some governments acting in the name of social justice.

Credit and behavioural scoring are some of the most important forecasting techniques used in the retail and consumer finance areas. As a pure forecasting tool as opposed to a decision-making one, credit scoring has mainly been used as a way of forecasting future bad debt in order to set aside appropriate provisioning. With the connections being made between scoring for default and scoring for targeting potential sales, these scoring techniques will clearly be used to forecast the sales of products as well as the profit a company will make in the future. There continue to be exciting developments and interesting problems waiting to be solved in this area and the changes in the capturing and storage of consumer data will give even more impetus to scoring methods.

That there will be progress in credit and behavioural scoring there can be no doubt. As the British author Samuel Butler said with uncanny forecasting ability nearly 100 years ago.

> *All progress is based upon a universal innate desire of every organism to live beyond its income.*

By that token, progress in credit scoring is a tautology.

Acknowledgements

We acknowledge the support of the Carnegie Trust which supported our travel to the University of Virginia where part of this research was undertaken. The authors is grateful to a referee for his careful reading and useful comments on an earlier version of the paper.

References

Aggarawal, A. (1990). Categorical data analysis, Wiley, New York.

Albright, H. T. (1994). Construction of a polynomial classifier for consumer loan applications using genetic algorithms, Department of Systems Engineering, University of Virginia, Working Paper.

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis; Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking and Finance 18*, 505–529.

Banasik, J., Crook, J. N., & Thomas, L. C. (1996). Does scoring a subpopulation make a difference? *International Review of Retail, Distribution and Consumer Research 6*, 180–195.

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when borrowers default. *Journal of Operational Research Society 50*, 1185–1190.

Bierman, H., & Hausman, W. H. (1970). The credit granting decision. *Management Science 16*, 519–532.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy 81*, 637–654.

Boyle, M., Crook, J. N., Hamilton, R., & Thomas, L. C. (1992). Methods for credit scoring applied to slow payers. In: Thomas, L. C., Crook, J. N., & Edelman, D. B. (Eds.), Credit scoring and credit control, Oxford University Press, Oxford, pp. 75–90.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees, Wadsworth, Belmont, CA.

Buckley, J., & James, I. (1979). Linear regression with censored data. *Biometrika 66*, 429–436.

Capon, N. (1982). Credit scoring systems: a critical analysis. *Journal of Marketing 46*, 82–91.

Carter, C., & Catlett, J. (1987). Assessing credit card applications using machine learning. *IEEE Expert 2*, 71–79.

Chandler, G. G., & Ewert, D. C. (1976). Discrimination on basis of sex and the Equal Credit Opportunity Act, Credit Research Centre, Purdue University, Indiana.

Chatterjee, S., & Barcun, S. (1970). A nonparametric approach to credit screening. *Journal of American Statistical Association 65*, 150–154.

Cheng, B., & Titterington, D. M. (1994). Neural networks: a review from a statistical perspective. *Statistical Science 9*, 2–30.

Choi, S. C. (1986). Statistical methods of discrimination and classification, Pergamon Press, New York.

Churchill, G. A., Nevin, J. R., & Watson, R. R. (1977). The role of credit scoring in the loan decision. *Credit World 3*(March), 6–10.

Coffman, J. Y. (1986). The proper role of tree analysis in forecasting the risk behaviour of borrowers, Management Decision Systems, Atlanta, MDS Reports 3,4,7, and 9.

Corcoran, A. W. (1978). The use of exponentially smoothed transition matrices to improve forecasting of cash flows from accounts receivable. *Management Science 24*, 732–739.

Crook, J. N. (2000). The demand for household debt in the US: evidence from the survey of consumer finance. Applied Financial Economics (in press).

Crook, J. N., Hamilton, R., & Thomas, L. C. (1992). The degradation of the scorecard over the business cycle. *IMA Journal of Mathematics Applied in Business and Industry 4*, 111–123.

Cyert, R. M., Davidson, H. J., & Thompson, G. L. (1962). Estimation of allowance for doubtful accounts by Markov chains. *Management Science 8*, 287–303.

Davis, D. B. (1987). Artificial Intelligence goes to work. *High Technology 4*(April), 16–17.

Davis, R. H., Edelman, D. B., & Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. *IMA Journal of Mathematics Applied in Business and Industry 4*, 43–52.

Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit environment. *European Journal of Operational Research 95*, 24–37.

Desai, V. S., Convay, D. G., Crook, J. N., & Overstreet, G. A. (1997). Credit scoring models in the credit union environment using neural networks and genetic algorithms. *IMA Journal of Mathematics Applied in Business and Industry 8*, 323–346.

Dirickx, Y. M. I., & Wakeman, L. (1976). An extension of the Bierman–Hausman Model for credit granting. *Management Science 22*, 1229–1237.

Durand, D. (1941). Risk elements in consumer installment financing, National Bureau of Economic Research, New York.

Edelman, D. B. (1997). Credit scoring for lending to small businesses. In: Proceedings of Credit Scoring and Credit Control V, Credit Research Centre, University of Edinburgh.

Eisenbeis, R. A. (1977). Pitfalls in the application of discriminant analysis in business, finance and economics. *Journal of Finance 32*, 875–900.

Eisenbeis, R. A. (1978). Problems in applying discriminant analysis in credit scoring models. *Journal of Banking and Finance 2*, 205–219.

ECOA (1975). Equal Credit Opportunity Act, U.S.C., Title 15, Sec. 1691 et seq.

ECOA (1976). Equal Credit Opportunity Act Amendments of 1976, US Government Printing Office, Washington, DC, Report of the Committee on Banking Housing and Urban Affairs, 94th Congress.

Fishelson-Holstine, H. (1998). Case studies in credit risk model development. In: Mays, E. (Ed.), Credit risk modeling, Glenlake Publishing, Chicago, pp. 169–180.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*, 179–188.

Fogarty, T. C., & Ireson, N. S. (1993). Evolving Bayesian

classifiers for credit control — a comparison with other machine learning methods. *IMA Journal of Mathematics Applied in Business and Industry 5*, 63–76.

Freed, N., & Glover, F. (1981a). A linear programming approach to the discriminant problem. *Decision Sciences 12*, 68–74.

Freed, N., & Glover, F. (1981b). Simple but powerful goal programming formulations for the discriminant problem. *European Journal of Operational Research 7*, 44–60.

Frydman, H., Kallberg, J. G., & Kao, D. -L. (1985). Testing the adequacy of Markov chains and Mover–Stayer models as representations of credit behaviour. *Operations Research 33*, 1203–1214.

Fung, R., Lucas, A., Oliver, R., & Shikaloff, N. (1997). Bayesian networks applied to credit scoring. In: Proceedings of Credit Scoring and Credit Control V, Credit Research Centre, University of Edinburgh.

Glen, J. J. (1997). Integer programming models for normalisation and variable selection in mathematical programming models for discriminant analysis. In: Proceedings of Credit Scoring and Credit Control V, Credit Research Centre, University of Edinburgh.

Grablowsky, B. J., & Talley, W. K. (1981). Probit and discriminant functions for classifying credit applicants; a comparison. *Journal of Economics and Business 33*, 254–261.

Hand, D. J. (1981). Discrimination and classification, Wiley, Chichester.

Hand, D. J., & Henley, W. E. (1993). Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry 5*, 45–55.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit. *Journal of the Royal Statistical Society, Series A 160*, 523–541.

Hand, D. J., & Jacka, S. D. (1998). Statistics in Finance, Arnold, London.

Hand, D. J., Oliver, J. J., & Lunn, A. D. (1996). Discriminant analysis when the classes arise from a continuum. *Pattern Recognition 31*, 641–650.

Hand, D. J., McConway, K. J., & Stanghellini, E. (1997). Graphical models of applications for credit. *IMA Journal of Mathematics Applied in Business and Industry 8*, 143–155.

Hardy, W. E., & Adrian, J. L. (1985). A linear programming alternative to discriminant analysis in credit scoring. *Abribus 1*, 285–292.

Henley, W.E. (1995). Statistical aspects of credit scoring. PhD thesis, Open University.

Henley, W. E., & Hand, D. J. (1996). A k-NN classifier for assessing consumer credit risk. *The Statistician 65*, 77–95.

Hopper, M. A., & Lewis, E. M. (1992). Behaviour scoring and adaptive control systems. In: Thomas, L. C., Crook, J. N., & Edelman, D. B. (Eds.), Credit scoring and credit control, Oxford University Press, Oxford, pp. 257–276.

Hsia, D. C. (1978). Credit scoring and the Equal Credit Opportunity Act. *The Hastings Law Journal 30*, 371–448.

Ignizio, J. P., & Soltys, J. R. (1996). An ontogenic neural network bankruptcy classification tool. *IMA Journal of Mathematics Applied in Business and Industry 7*, 313–326.

Joachimsthaler, E. A., & Stam, A. (1990). Mathematical programming approaches for the classification problem in two-group discriminant analysis. *Multivariate Behavioural Research 25*, 427–454.

Joanes, D. N. (1993). Reject inference applied to logistic regression for credit scoring. *IMA Journal of Mathematics Applied in Business and Industry 5*, 35–43.

Johnson, R. W. (1992). Legal, social and economic issues implementing scoring in the US. In: Thomas, L. C., Crook, J. N., & Edelman, D. B. (Eds.), Credit scoring and credit control, Oxford University Press, Oxford, pp. 19–32.

Jost, A. (1998). Data mining. In: Mays, E. (Ed.), Credit risk modeling, Glenlake Publishing, Chicago, pp. 129–154.

Kolesar, P., & Showers, J. L. (1985). A robust credit screening model using categorical data. *Management Science 31*, 123–133.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association 70*, 782–790.

Lachenbruch, P. A. (1975). Discriminant analysis, Hafner Press, New York.

Lai, T. L., & Ying, Z. L. (1994). A Missing information principle and M-estimators in regression analysis with censored and truncated data. *Annals of Statistics 22*, 1222–1255.

Lando, D. (1994). Three essays on contingent claims pricing. PhD thesis, Cornell University, Ithaca.

Leonard, K. J. (1993a). Empirical Bayes analysis of the commercial loan evaluation process. *Statistics and Probability Letters 18*, 289–296.

Leonard, K. J. (1993b). Detecting credit card fraud using expert systems. *Computers and Industrial Engineering 25*, 103–106.

Lewis, E. M. (1992). An introduction to credit scoring, Athena Press, San Rafael, CA.

Li, H. G., & Hand, D. J. (1997). Direct versus indirect credit scoring classification. In: Proceedings of Credit Scoring and Credit Control V, Credit Research Centre, University of Edinburgh.

Lovie, A. D., & Lovie, P. (1986). The flat maximum effect and linear scoring models for prediction. *Journal of Forecasting 5*, 159–186.

Makowski, P. (1985). Credit scoring branches out. *The Credit World 75*, 30–37.

Mangasarian, O. L. (1965). Linear and nonlinear separation of patterns by linear programming. *Operations Research 13*, 444–452.

Mangasarian, O. L. (1993). Mathematical Programming in Neural Networks. *ORSA Journal on Computing 5*, 349–360.

Martell, T. F., & Fitts, R. L. (1981). A quadratic discriminant analysis of bank credit card user characteristics. *Journal of Economics and Business 33*, 153–159.

Mays, E. (1998). Credit risk modeling, Glenlake Publishing, Chicago.

Mehta, D. (1968). The formulation of credit policy models. *Management Science 15*, 30–50.

Merton, R. C. (1974). On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance 29*, 449–470.

Myers, J. H., & Forgy, E. W. (1963). The development of numerical credit evaluation systems. *Journal of American Statistics Association 58*(September), 799–806.

Narain, B. (1992). Survival analysis and the credit granting decision. In: Thomas, L. C., Crook, J. N., & Edelman, D. B. (Eds.), Credit scoring and credit control, Oxford University Press, Oxford, pp. 109–122.

Nath, R., Jackson, W. M., & Jones, T. W. (1992). A comparison of the classical and the linear programming approaches to the classification problem in discriminant analysis. *Journal of Statistical Computation and Simulation 41*, 73–93.

Oliver, R. M. (1993). Effects of calibration and discrimination on profitability scoring. In: Proceedings of Credit Scoring and Credit Control III, Credit Research Centre, University of Edinburgh.

Orgler, Y. E. (1971). Evaluation of bank consumer loans with credit scoring models. *Journal of Bank Research 1*(Spring), 31–37.

Overstreet, G. A., Bradley, E. L., & Kemp, R. S. (1992). The flat maximum effect and generic linear scoring models: a test. *IMA Journal of Mathematics Applied in Business and Industry 4*, 97–110.

Platts, G., & Howe, I. (1997). A single European scorecard. In: Proceedings of Credit Scoring and Credit Control V, Credit Research Centre, University of Edinburgh.

Quinlan, J. R. (1993). C4.5: programs for machine learning, Morgan Kaufman, San Mateo, CA.

Reichert, A. K., Cho, C. -C., & Wagner, G. M. (1983). An examination of the conceptual issues involved in developing credit scoring models. *Journal of Business and Economic Statistics 1*, 101–114.

Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: a survey. *Operations Research 42*, 589–613.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man and Cybernetics 21*, 660–674.

Sewart, P., & Whittaker, J. (1998). Fitting graphical models to credit scoring data. *IMA Journal of Mathematics in Business and Industry 9*, 241–266.

Showers, J. L., & Chakrin, L. M. (1981). Reducing revenue from residential telephone customers. *Interfaces 11*, 21–31.

Srinivasan, V., & Kim, Y. H. (1987a). The Bierman–Hausman credit granting model: a note. *Management Science 33*, 1361–1362.

Srinivasan, V., & Kim, Y. H. (1987b). Credit granting: a comparative analysis of classification procedures. *Journal of Finance 42*, 665–683.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. *Management Science 38*, 926–947.

Tessmer, A. C. (1997). What to learn from near misses: on inductive learning approach to credit risk assessment. *Decision Sciences 28*, 105–120.

Thomas, L. C. (1992). Financial risk management models. In: Ansell, J., & Wharton, F. (Eds.), Risk analysis, assessment and management, Wiley, Chichester.

Thomas, L. C. (1994). Applications and solution algorithms for dynamic programming. *Bulletin of the IMA 30*, 116–122.

Thomas, L. C. (1998). Methodologies for classifying applicants for credit. In: Hand, D. J., & Jacka, S. D. (Eds.), Statistics in finance, Arnold, London, pp. 83–103,.

Thomas, L. C., Crook, J. N., & Edelman, D. B. (1992). Credit scoring and credit control, Oxford University Press, Oxford.

Thomas, L. C., Allen, D., & Morkel-Kingsbury, N. (1998). A hidden Markov Chain model of credit risk spreads, Department of Finance and Business Economics, Edith Cowan University, Working Paper.

Titterington, D. M. (1992). Discriminant analysis and related topics. In: Thomas, L. C., Crook, J. N., & Edelman, D. B. (Eds.), Credit scoring and credit control, Oxford University Press, Oxford, pp. 53–73.

Vlachonikolis, I. G. (1986). On the estimation of the expected probability of misclassification in discriminant analysis with mixed binary and continuous variables. *Computers and Maths with Applications 12A*, 187–195.

van Kuelen, J. A. M., Spronk, J., & Corcoran, A. W.

(1981). Note on the Cyert–Davidson–Thompson doubtful accounts model. *Management Science 27*, 108–112.

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behaviour. *Journal of Financial and Quantitative Analysis 15*, 757–770.

Yobas, M. B., Crook, J. N., & Ross, P. (1997). Credit scoring using neural and evolutionary techniques, Credit Research Centre, University of Edinburgh, Working Paper 97/2.

Zandi, M. (1998). Incorporating economic information into credit risk underwriting. In: Mays, E. (Ed.), Credit risk modeling, Glenlake Publishing, Chicago, pp. 155–168.

Ziari, H. A., Leatham, D. J., & Ellinger, P. N. (1997). Development of statistical discriminant mathematical programming model via resampling estimation techniques. *American Journal of Agricultural Economics 79*, 1352–1362.

Zocco, D. P. (1985). A framework for expert systems in bank loan management. *Journal of Commercial Bank Lending 67*, 47–54.