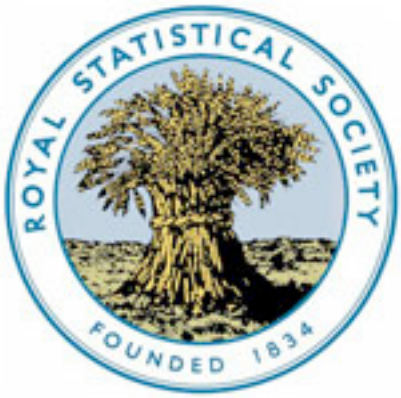

J. R. Statist. Soc. A (1997)
160, *Part* 3, *pp.* 523–541

Statistical Classification Methods in Consumer Credit Scoring: a Review

By D. J. HAND† and W. E. HENLEY

The Open University, Milton Keynes, UK *Rothschilds, London, UK*

[Received March 1995. Final revision October 1996]

SUMMARY

Credit scoring is the term used to describe formal statistical methods used for classifying applicants for credit into 'good' and 'bad' risk classes. Such methods have become increasingly important with the dramatic growth in consumer credit in recent years. A wide range of statistical methods has been applied, though the literature available to the public is limited for reasons of commercial confidentiality. Particular problems arising in the credit scoring context are examined and the statistical methods which have been applied are reviewed.

Keywords: CLASSIFICATION; CONSUMER LOANS; CREDIT CONTROL; CREDIT SCORING; DISCRIMINANT ANALYSIS; FINANCE; REJECT INFERENCE; RISK ASSESSMENT

1. INTRODUCTION

In this paper we use the term 'credit' to refer to an amount of money that is loaned to a consumer by a financial institution and which must be repaid, with interest, in instalments (usually at regular intervals). We focus chiefly on methods for classifying an applicant for credit into classes according to their likely repayment behaviour (e.g. 'default' or 'not default' with repayments), but we shall also briefly consider other associated problems in the credit industry. The probability that an applicant will default must be estimated from information about the applicant provided at the time of the application, and the estimate will serve as the basis for an accept or reject decision. Accurate classification is of benefit both to the creditor (in terms of increased profit or reduced loss) and to the applicant (avoiding overcommitment).

This activity—deciding whether or not to grant credit—is carried out by banks, building societies, retailers, mail order companies, utilities and various other organizations. It is an economic activity which has seen rapid growth over the last 30 years. Some figures will illustrate the size of the consumer credit industry. The total UK consumer debt, including mortgages, bank loans, debts to retailers, credit card debts, etc., is about £500 billion. In 1994 in the UK about 12% of retail expenditure was made using credit cards, amounting to a total of about £36 billion. Credit card spending increased by about 16% between January 1995 and January 1996. Around 10 million British households currently have mortgages.

Traditional methods of deciding whether to grant credit to a particular individual use human judgment of the risk of default, based on experience of previous decisions. However, economic pressures resulting from increased demand for credit, allied with

†*Address for correspondence*: Department of Statistics, Faculty of Mathematics and Computing, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.
E-mail: d.j.hand@open.ac.uk

 0964–1998/97/160523

greater commercial competition and the emergence of new computer technology, have led to the development of sophisticated statistical models to aid the credit granting decision. *Credit scoring* is the name used to describe this more formal process of determining how likely applicants are to default with their repayments. (Sometimes the term *application scoring* is used to distinguish it from *behavioural* or *performance scoring*, which refers to monitoring and predicting the repayment behaviour of a consumer to whom credit has already been granted.) Statistical models, called *score-cards* or *classifiers*, use predictor variables from application forms and other sources to yield estimates of the probabilities of defaulting. An accept or reject decision is taken by comparing the estimated probability of defaulting with a suitable threshold. Standard statistical methods used in the industry for developing score-cards are discriminant analysis, linear regression, logistic regression and decision trees. In the industry, the predictor variables are typically called *characteristics*, a terminology which we shall retain here. The values that they can take are called *attributes*.

Loans may be fixed term—the repayment calculations are such that the loan and interest will be totally repaid after a certain time—or they may be *rolling* or *revolving* loans, such as with credit cards, where the loan amount can be increased flexibly. The length of a fixed term loan repayment period will, of course, depend on the nature of the loan. A mortgage may be over a 25-year period, whereas a car loan will be for a much shorter period, and the repayment period for catalogue purchases shorter still. One might be interested in the probability that the borrower will have defaulted by the end of the loan period or in more subtle measures, such as the probability that they are two or three payments behind at the end of 1 year.

One complication is that the population of potential borrowers evolves owing to selection processes. Hand *et al.* (1996a) have described how an initial population goes through various selection steps: the bank decides to whom to mail an application form; only some of these return the form; the bank scores these and offers some a loan; only some of these accept the loan; of these some default and some repay early; and finally the bank decides to whom to offer a further loan.

The data from which a score-card will be constructed will be that in a (*design*) sample of applicants to whom credit has already been granted. Normally this will include the values of their characteristics (which here we take to include information from *credit bureaux*—see Section 2) and also the true class (which, for brevity, we shall here call 'good' or 'bad' risk classes) to which they were found to belong. Thus, implicit in all the work that we discuss in this paper, we make a straightforward division of applicants into two groups. More sophisticated approaches can categorize the design sample applicants on a (possibly multidimensional) continuum according to their repayment behaviour, and this more refined scale can be used as the response variable(s) in the model. Similarly, the fact that people are subject to influences which can change their propensity to default (external social or financial pressures, for example) will also not be discussed here; such things could be integrated into a behavioural scoring model. Indeed, although default risk is the focus of the greater part of credit scoring effort in the industry, one can argue that this risk is only one aspect of the overall credit granting decision. The main aim will normally be to maximize profits, and profitability need not be monotonically related to risk. For example, very low risk applicants who pay off their credit card bill each month, so that the lender cannot charge interest, are not profitable. Conversely, very high risk

applicants can be profitable, provided that a sufficiently high rate of interest is charged. In general, devising suitable operational definitions of profitability which capture the important aspects of the idea is not straightforward. Factors which need to be considered are the cost of collecting and analysing information (and hence the possibility of the multistage approaches mentioned in Section 2), expected returns on good and bad loans that have been accepted, the fact that loans may be profitable even if the borrowers default, the *attrition* rate (the probability that someone accepted for a loan will decline the offer—which clearly depends on the broader competitive environment) and factors such as interest rates. Some work in this direction is described in Greer (1967), Edmister and Schlarbaum (1974), Oliver (1992) and Hand *et al.* (1996a).

Population drift can also be a problem in credit scoring applications. This describes the tendency of populations to evolve with time, so that the distributions change. This is to be expected since applicant populations are subject to economic pressures and a changing competitive environment. Population drift is different from changes in the population induced by selection steps in the loan process, mentioned above. The effect of population drift on a score-card will be to degrade its performance—although the distributions of scores tend to change only slowly compared with the level of acceptable risk. The latter can be handled by adjusting the classification thresholds, whereas the former requires the periodic production of new score-cards (or the use of a method which permits dynamic updating). Sometimes population drift can be expected *a priori*, so serving as a stimulus for the production of a new score-card (e.g. the differences before and after a general election). In general, however, without such obvious indications that a new score-card may be needed, it is necessary to look for more subtle signs. Standard approaches are to compare the statistical properties of the applicant population (at various times) with those of the development sample. This can be done on the basis of overall score or on the basis of individual characteristics or credit bureau items.

A practice which has been common in the credit industry is to define three classes of risk (good and bad risk as above and indeterminate), to design the score-card using only the extreme two classes, and then to classify new applicants as either good or bad risks. (This is distinct from classifying new applicants as 'bad', 'good' or 'not yet known'—and then seeking further information on the last group.) For example, good risks might be those borrowers who have never been in arrears, bad risks might be those who have been three or more consecutive repayments in arrears at some point during the period in question and indeterminates might be those who have been in arrears either for one or for two consecutive repayments. This practice seems curious and difficult to justify. We speculate that the practice arose because 'default' is a woolly concept. What the lender is really interested in is whether or not the customer will yield a profit. Those defined as good risk according to a 'never in arrears' definition might definitely yield a profit, those defined as bad risk according to the 'three months in arrears' might definitely not be profitable and the indeterminates might or might not be profitable—depending on such unpredictables as economic changes over the term of the loan. The standard method then seeks to construct a rule which separates the definitely profitable from the definitely unprofitable—a perfectly reasonable objective.

Another problem of particular relevance in credit scoring is that, in general, only those who are accepted for credit will be followed up to find out whether they do turn

out to be good or bad risks according to the definition adopted, so that the design sample will be a biased sample from the overall population of applicants. Attempts to tackle this, using what information there is on the rejected applicants (namely their values on the characteristics, but not their true classes) are called *reject inference*, and we discuss these below. These issues are distinct from those arising due to *ineligibles* and *overrides*. The former describes those who lie outside the scope of the system. The latter describes a decision by the management which differs from that produced by the scoring system. This may be the rejection of an applicant whose score is on the accept side of the threshold or the converse. Overrides may occur because of extra information available to management or because of policy rules that applicants with certain characteristic values are to be treated in a predetermined way. In general they will not lead to biased samples, as above, if the relevant applicant population is defined exclusively of those eliminated by a high side override.

Despite the widespread use of what are essentially statistical techniques in the consumer credit industry, the published literature seems relatively sparse. The main reason for this seems to be the need for confidentiality: superior techniques provide a competitive advantage, so that an organization will not be too keen to divulge them. Moreover, data sets containing confidential information on applicants cannot be released to third parties without careful safeguards. However, it is apparent that industry practice does not always reflect what academic statisticians might regard as the best approaches.

2. THE DATA

Credit scoring databases are often large: well over 100000 applicants measured on more than 100 variables are quite common. (Databases for behavioural scoring, containing information about past repayment behaviour, can be much larger.) The proportion of applicants to whom credit is extended varies greatly—we have worked on examples where the proportion is as low as 17% and examples where it is as high as 84%. This proportion will depend on the financial product in question, the target population and the risk that the lenders are willing to assume. In principle, the risk can be made arbitrarily small by selecting only those applicants (a proportion of 5% is quite common) who are thought to have a very low risk of defaulting. Thus the proportion accepted and the proportion of those accepted who subsequently default are inversely related. In some situations, such as mail order purchasing from a catalogue, each accepted applicant will lead to a cost (due to the printing and distribution of the catalogue), whether or not they subsequently turn out to be a good risk. If the number of good risks substantially outweighs the number of bad risks, then this initial cost may be the greatest component of cost. Because of this it is common in such situations to fix the number of applicants to be accepted beforehand. A figure of 70% accepted is quite usual.

Table 1 shows the sorts of characteristics that are used in credit scoring. Naturally these will vary from situation to situation: someone seeking to purchase from a mass home shopping catalogue will be asked for different information from someone seeking a £500000 mortgage. Moreover, the information which may be used in a credit scoring system is subject to changing legislation. Currently in the UK, for example, it is not permissible to discriminate on gender grounds.

In some cases the values of the characteristics are obtained sequentially: an initial

TABLE 1
Characteristics typical of certain credit scoring domains

Time at present address	0–1, 1–2, 3–4, 5+ years
Home status	Owner, tenant, other
Postcode	Band A, B, C, D, E
Telephone	Yes, no
Applicant's annual income	£(0–10000), £(11000–20000), £(21000+)
Credit card	Yes, no
Type of bank account	Cheque and/or savings, none
Age	18–25, 26–40, 41–55, 55+ years
County Court judgments	Number
Type of occupation	Coded
Purpose of loan	Coded
Marital status	Married, divorced, single, widow, other
Time with bank	Years
Time with employer	Years

screening based on an application form identifies unequivocal good and bad risk applicants, and these are accepted or rejected immediately. Further information is then sought on the remainder, either by means of further forms or from credit reference agencies. The sequential process avoids the costs of obtaining unnecessary information (the costs to the applicant—in terms of time and irritation about having to complete yet further forms—or the cost to the lender due to the charges made by the credit reference agency) as well as permitting a quick decision for the majority of the applicants. This latter point can be important: a lengthy process is likely to deter those who do not really need the loan or have other sources of credit, i.e. the good risks may be deterred from requesting a loan.

Credit reference agencies collect information on the past behaviour of applicants—information such as the number and details of loan accounts, details of slow payments, bankruptcies, the number of requests for new credit and so on. Information about almost every adult in the UK is kept on such databases.

As can be seen from Table 1, the data are often categorical (typically, continuous variables are categorized), usually with only a few categories, though some, such as postcode, can have many categories. Although the range of statistical techniques for handling multivariate categorical data has widened dramatically in the last 15 years, almost all commercial credit scoring systems use dummy variables (see, for example, Crook *et al.* (1992)). However, the alternative approach of coding categorical variables into numeric form and using continuous data models is becoming more common. For example, one strategy is to use logarithms of likelihood ratios, in this context called *weights of evidence*: the jth attribute of the ith characteristic is scored as

$$w_{ij} = \ln(p_{ij}/q_{ij}),$$

where p_{ij} is the number of good risks in attribute j of characteristic i divided by the total number of good risks (who respond to characteristic i) and q_{ij} is the number of bad risks in attribute j of characteristic i divided by the total number of bad risks (who respond to characteristic i). An alternative approach is to use optimal scaling (Gifi, 1990).

Data for credit scoring usually have the common characteristic of multivariate data that there are missing values. Such values may be structurally missing (e.g. questions which are only asked conditionally on the responses to previous questions) or randomly missing. In one of our studies there were 3883 applicants in the design set with values recorded for 25 characteristics. Of the 3883 applicants, only 66 had no missing values and one had 16 missing values. Of the 25 characteristics, just five had no missing values and two had over 2000 missing values. Strategies for coping with missing components of measurement vectors in discrimination problems have been developed by statisticians. They include coding a missing value as an additional attribute, dropping incomplete vectors (from the design set), substituting values for the missing values, substituting values iteratively in conjunction with a model (as in the EM algorithm, for example), or carrying out the classification in the appropriate marginal space. Sometimes the first of these strategies can yield usefully discriminating information: a refusal to answer a particular question may be indicative of greater risk.

As in many classification problems, there are complementary pressures on the number of variables to be included. Since the data sets are generally large, overfitting problems may not occur. Thus one might seek to use as many variables as possible. However, there are practical limitations: as mentioned above, too many questions or too lengthy a vetting procedure will deter applicants, who will go elsewhere. A standard statistical and pattern recognition strategy here is to explore a large number of characteristics (in credit scoring 50 or more are quite common: Duffy (1977) refers to 300 and Scallan (personal communication) refers to a case with 2500) for the design sample, and to identify an effective subset (of say 10–12) of those characteristics for application in practice.

In credit scoring three approaches to selecting characteristics are commonly used.

(a) Using expert knowledge, experience and a feeling for the data and characteristics provides a good complement to the formal statistical manipulations. The latter will prevent unpredictive characteristics being included for historical reasons whereas the former will be essential if asked to justify the chosen selection of characteristics. It is necessary (if unfortunate from a purist's point of view) to be able to justify the system to non-statistical users: too complex a system will be unacceptable, even if it outperforms simpler systems. This manifests itself, for example, in *inversions*: the users may expect to see a monotonic increase in risk with the ordered attributes of some characteristic and may prefer to avoid using a score-card where the relationship is non-monotonic. For example, Capon (1982) described a scoring table produced for the finance subsidiary of a consumer durables manufacturer in which the relationship between monthly income and points to be added to the overall score is not monotonic.

(b) Using stepwise statistical procedures is the second approach. For example, forward stepwise methods sequentially add variables, at each step adding that variable (or group of variables) which lead(s) to the greatest improvement in predictive accuracy. Problems such as the inversions mentioned above can arise if stepwise methods are used with dummy variables—since then perhaps only certain categories of a variable will be selected.

(c) The third approach is to select individual characteristics by using a measure of

the difference between the distributions of the good and bad risks on that characteristic. One common such measure is the *information value*, defined as

$$\sum_j (p_{ij} - q_{ij})\, w_{ij}$$

where p_{ij}, q_{ij} and w_{ij} are as above. Typically any characteristic with an information value of over 0.1 will be considered for inclusion in the score-card. Another common measure is the χ^2-statistic derived from a cross-tabulation of class (good or bad risk) by the attributes of the characteristic in question. From the perspective of multivariate statistics, such an approach has obvious shortcomings.

In practice these methods will typically all be used, perhaps beginning with an initial selection on an individual basis, eliminating using stepwise methods, adding understanding of the domain and so on, in a sequential and iterative manner.

3. ASSESSMENT OF PERFORMANCE

Because of the large data sets that are available, validation can usually be based on a test set—complications such as bootstrap or jackknife methods do not normally need to be considered. There are basically two classes of assessment method: separability measures of the good and bad risks' score distributions and counting methods.

Given that each applicant is assigned a score (by linear regression or one of the other methods described in Section 4), common separability measures used in this context are the *divergence statistic* (the value of the sample t-statistic between the two design set classes) and the *information statistic* (defined like the information value above, but for the distribution over scores rather than over a characteristic). Wilkie (1992) and Hand (1994) have reviewed such measures.

Counting measures are based on the 2×2 table of predicted-by-true classes, i.e. a threshold is imposed on the scores such that applicants scoring below the threshold are predicted to be bad risks and those above to be good risks. A *Lorentz diagram* is sometimes used, showing the curve of the cumulative proportion of true good risks plotted against the cumulative proportion of true bad risks as the threshold varies. (An example is given in Fig. 1.) This is equivalent to *receiver–operating characteristic analysis* (Zweig and Campbell, 1993) in which (usually) the *true positive rate* (the proportion of the true good risks who are above the threshold) is plotted against the *false positive rate* (the proportion of true bad risks who are above the threshold). An ideal classifier in such a plot would follow the axes, and the area between the curve and the axes (or some equivalent transformation, e.g. the *Gini coefficient*, which is twice the area between the curve and the diagonal) is sometimes used as a measure of the discriminatory power of the score-card. This measure provides a global summary of performance, integrated in some sense over all possible choices of threshold. This is reasonable since, as commented in Section 1, the threshold levels are often unstable compared with the rank ordering of estimated probabilities.

A 2×2 table has four numbers in it. If the total is fixed and the proportion of good risks in the population (or test sample) is fixed, then we have 2 degrees of

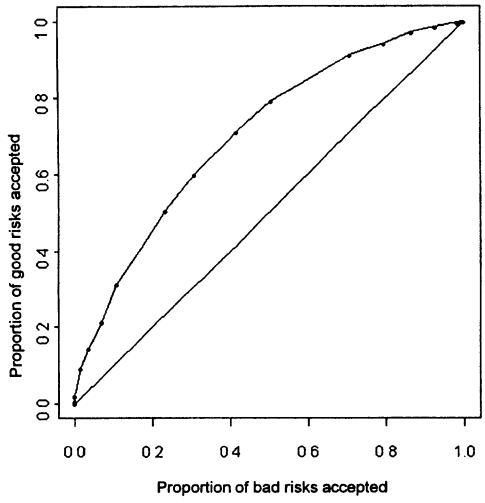

Fig. 1. Example of a Lorentz diagram, for a nearest neighbour classifier (Section 4.2.8) applied to a set of loans for home improvements: as the threshold varies, the curve shows the proportions of the good risks who are accepted (vertical axis) plotted against the proportion of bad risks who are accepted (horizontal axis); a perfect score-card would follow the left vertical and top horizontal axes, accepting 100% of the good risks before accepting any of the bad risks; a score-card which performed randomly at all thresholds would follow the diagonal line

freedom. To produce a uniquely 'best' classifier this must somehow be reduced to 1. Sometimes (as with other classification problems) the error rate is used. This simply regards each type of misclassification as equally serious and counts the total number misclassified. An alternative, which is common in some types of credit domain, is to fix 1 degree of freedom. For example, as mentioned earlier, in mail order each accepted applicant incurs a significant cost (supplying them with a catalogue) whether they turn out to be good or bad risks. Because of this, the proportion to be accepted is fixed beforehand, leaving 1 degree of freedom, which can be expressed as the bad risk rate among those accepted. This criterion has the property that it is bounded by numbers which may be greater than 0 and less than 1. For example, given that a proportion p of the population are good risks and that a proportion a are to be accepted, with $a > p$ then the bad risk rate among those accepted can be no less than $1 - p/a$. Henley and Hand (1996) have discussed such bounds in detail. For the data set described in that study $p = 0.45$ and $a = 0.70$, leading to a bad risk rate that necessarily lies between 0.35 and 0.78. Such bounds permit us to put score-card performances into context: a bad risk rate far from 0 does not necessarily mean that the score-card is poor.

4. METHODS OF IDENTIFYING GOOD AND BAD RISKS

4.1. *Judgmental versus Scoring Methods*

Before formal scoring methods became widespread, the credit granting decision was based on subjective human assessment. Quite naturally, the introduction of statistical methods encountered some scepticism. Nowadays this has been overcome: not only are scoring methods the only way of handling the large number of

transactions, but it seems that they produce more accurate classifications than subjective *judgmental* assessments by human experts (Rosenberg and Gleit, 1994). This will be no surprise to those with experience in the analogous areas of clinical reasoning and diagnosis. For example, McGuire (1985) says:

> '. . . studies of clinical reasoning all too often have revealed disquieting defects in the process, namely, that physicians often fail to collect the data they need, to pay attention to the data they do collect, to use their knowledge effectively in making interpretations of and inferences from the data they do consider, to incorporate a systematic consideration of alternative risks and values in the actions they take on the basis of those inferences . . .'.

These issues are discussed further in Hand (1985), chapter 5.

Chandler and Coffman (1979) have also compared classifications based on the experience and judgment of a human assessor with those based on statistical scoring processes. They pointed out that simple accuracy of creditworthiness predictions is not the only thing which must be taken into account. Other factors to consider are managerial ability to exercise effective control over the credit granting process, the ability to forecast results and to prepare relevant management reports, compliance with legal constraints and social and political acceptability. They conclude:

> '. . . it seems that, on the whole, the empirical evaluation process has no serious deficiencies not also shared by judgemental evaluation. It also appears that empirical evaluation of creditworthiness has certain advantages that do not exist with judgemental evaluation. On the other hand, judgemental evaluation may have an advantage in dealing with individual cases that truly are exceptions from past experience.'

Similarly, Reichert *et al.* (1983), although not convinced of the predictive ability of scoring approaches, observed that (p. 102)

> 'their real benefit may relate not to any superiority in predictive power but to the highly consistent, objective, and efficient manner in which such predictions are made'.

They might also have added the fact that scoring is typically cheaper than the alternative. Hsia (1978) has also discussed the disadvantages of judgmental systems.

Nowadays it seems that the only organizations which do not use credit scoring approaches are the smaller and/or more personal companies, and those concerned with corporate finance, where statistical methods have been slower to be adopted. However, although the financial community may have confidence in objective statistical credit scoring methods, there seems still to be some suspicion of them in the customer base. This stems in part from anxiety about the impersonal nature of the process and in part from concerns over the accuracy of the data relating to the individual applicant.

4.2. *Statistical Scoring Methods used in Practice*

Historically, discriminant analysis and linear regression have been the most widely used techniques for building score-cards. Both have the merits of being conceptually straightforward and widely available in statistical software packages. Typically the coefficients and the numerical scores of the attributes are combined to give single contributions which are added to give an overall score. Usually, these contributions are manipulated so that they are integral. Other techniques which have been used in the industry include logistic regression, probit analysis, nonparametric smoothing

methods, mathematical programming, Markov chain models, recursive partitioning, expert systems, genetic algorithms, neural networks and conditional independence models. If only a few characteristics are involved, with a sufficiently small number of attributes, an explicit classification table can be drawn up, showing the classification to be given to each combination of attributes.

In what follows we summarize these various methods, giving examples from the credit scoring literature describing their use. Section 4.3 assesses the relative strengths and weaknesses of the methods.

4.2.1. *Discriminant analysis*

The first published account of the use of discriminant analysis to produce a scoring system seems to be that of Durand (1941) who showed that the method could produce good predictions of credit repayment. Eisenbeis (1977, 1978) presented a critical assessment of the use of discriminant analysis in business, finance and economics in general. The criticisms are discussed in Rosenberg and Gleit (1994). In our view, the demerits of discriminant analysis have been overstressed in these papers. For example, Eisenbeis (1977), p. 213, said

> 'one of the critical assumptions in discriminant analysis is that the variables describing the members of the groups being evaluated are multivariate normally distributed'.

This is a common misconception. Certainly, if the variables follow a multivariate ellipsoidal distribution (of which the normal distribution is a special case), then the linear discriminant rule is optimal (ignoring sampling variation). However, if discriminant analysis is regarded as yielding that linear combination of the variables which maximizes a particular separability criterion, then clearly it is widely applicable. The normality assumption only becomes important if significance tests are to be undertaken. Eisenbeis (1978) also argued that discriminant analysis procedures are only legitimate when the 'groups being investigated are discrete and identifiable' and not, for example, 'when an inherently continuous variable is segmented and used as a basis to form groups'. However, Hand *et al.* (1996b) showed that the discriminant function obtained by segmenting a multivariate normal distribution into two classes is parallel to the optimal discriminant function, so this is not necessarily true. Arguments such as these help to explain why Reichert *et al.* (1983), on the basis of empirical observation of credit scoring problems, concluded that

> 'the fact that a significant portion of credit information is not normally distributed may not be a critical limitation'.

Other accounts of the use of discriminant analysis in credit scoring are given by Myers and Forgy (1963) (who compared discriminant analysis and regression analysis), Lane (1972), Apilado *et al.* (1974) and Moses and Liao (1987). Grablowsky and Talley (1981) compared linear discriminant analysis and probit analysis by using data from a large midwestern retail chain in the USA.

4.2.2. *Regression*

Ordinary linear regression has also been used for the two-class problem in credit scoring. Since regression using dummy variables for the class labels yields a linear

combination of the predicting characteristics which is parallel to the discriminant function (Lachenbruch, 1975), we might also expect this method to perform reasonably. Orgler (1970) used regression analysis in a model for commercial loans, and Orgler (1971) used regression analysis to construct a score-card for evaluating outstanding loans, rather than screening new applications. Since the evaluation of outstanding loans includes information about how the customer has performed so far, it is a behavioural scoring model. He found that the behavioural characteristics were more predictive of future loan quality than are the application characteristics. Other studies describing the use of regression include Fitzpatrick (1976), Lucas (1992) and Henley (1995).

4.2.3. *Logistic regression*

On theoretical grounds we might suppose that logistic regression is a more appropriate statistical tool than linear regression, given that two discrete classes (good and bad risks) have been defined. In a comparative study, however, Henley (1995) found that logistic regression was no better than linear regression. He attributed this to the fact that a relatively large proportion of the applicants whom he studied had scores associated with estimated probabilities of being good risks between 0.2 and 0.8. When this is the case the logistic curve is very well approximated by a straight line.

Wiginton (1980) gave one of the first published accounts of logistic regression applied to credit scoring, in a comparison with discriminant analysis. He concluded that the logistic approach gave superior classification results but that neither method was sufficiently good to be cost effective for his problem. However, his problem was unusual in that, from the eight characteristics available, only three were selected as significantly related to credit rating, and only these were used in the subsequent analysis, which used an indicator variable approach. Srinivasan and Kim (1987a) included logistic regression in a comparative study with other methods—although for a corporate credit granting problem. Leonard (1993a) also applied logistic regression to a commercial loan evaluation process (exploring several models, including a model using random effects for bank branches).

4.2.4. *Mathematical programming methods*

Given an objective criterion to optimize (such as the proportion of applicants correctly classified) we can cast the problem into a mathematical programming framework. For example, Hand (1981), chapter 4, described how to minimize the perceptron criterion (a linear function of the sum of the points corresponding to applicants who are misclassified) by using linear programming and Showers and Chakrin (1981) and Kolesar and Showers (1985) used integer programming to determine whether telephone customers should be required to leave a deposit. Mathematical programming techniques have the additional advantage that deterministic relationships between the characteristics pose no problems. This is not so with all competing methods; for example, a linear relationship between characteristics would lead to a singular covariance matrix for the predictors, so unmodified discriminant analysis (Section 4.2.1) could not be applied.

4.2.5. *Recursive partitioning*

Recursive partitioning or decision tree methods have been developed in several disciplines, most notably the life sciences, statistics and artificial intelligence. One of the most important references in statistics is Breiman *et al.* (1984) and a recent survey which also covers work in artificial intelligence is that of Safavian and Landgrebe (1991). Applications of such methods in credit scoring are described by Makowski (1985), Coffman (1986), Carter and Catlett (1987) and Mehta (1968). The last developed a partitioning method aimed at minimizing cost. Boyle *et al.* (1992) compared recursive partitioning with discriminant analysis. In fact, decision trees also occur in disguise in other methods of credit scoring. A *derogatory tree* is a small classification tree which can help to identify poor risk applicants and which is included in a score-card as if it were a single characteristic. Thus non-linearities and interactions between characteristics can be included in what is superficially a simple linear model.

4.2.6. *Expert systems*

Quite naturally, any new technology which shows potential for improved accuracy in predicting poor risk applicants will attract interest in a commercially competitive environment. With expert system shells now readily available, it is hardly surprising that they have been applied to credit scoring. Unfortunately published accounts are relatively rare and do not go into great detail. They include Zocco (1985), Davis (1987) and Leonard (1993b, c). One of the attractive features of expert systems in this application is the emphasis placed on their ability to explain their recommendations and decisions. This can be particularly important given the legal requirements for credit scorers to give reasons for rejecting applicants. Unfortunately, what limited evidence is available has suggested quite poor predictive performance for such approaches.

4.2.7. *Neural networks*

The type of neural network that is normally applied to credit scoring problems can be viewed as a statistical model involving linear combinations of nested sequences of non-linear transformations of linear combinations of variables. Other classes of statistical model have the same sort of flexibility (see, for example, Ripley (1994)). Rosenberg and Gleit (1994) described several applications of neural networks to corporate credit decisions and fraud detection and Davis *et al.* (1992) compared such methods with alternative classifiers.

4.2.8. *Smoothing nonparametric methods*

Nonparametric methods, especially nearest neighbour methods, have been explored for credit scoring applications, e.g. by Chatterjee and Barcun (1970), Hand (1986) and Henley and Hand (1996). The first of these studied personal loan applications to a New York bank and classified them on the basis of the proportion of cases with identical characteristic vectors which belonged to the same class (this is feasible since they had only eight characteristics, all binary). Hand (1986) compared a variety of classification methods, including nearest neighbours and recursive partitioning classifiers, on a data set describing applications for loans for home

improvement. Henley and Hand (1996) described a detailed investigation of nearest neighbour methods applied to data from a large mail order company. In particular, they investigated the choice of metric (how to define 'nearest') and the choice of the number of nearest neighbours to consider.

The nearest neighbour method has some attractive features for credit scoring applications. For example, it is straightforward to update dynamically by adding applicants to the design set when their true class becomes known and by dropping older cases, i.e. it can be used to overcome population drift. Despite this merit, nearest neighbour methods have not been widely adopted in the credit scoring industry. One reason for this is the perceived computational demand: not only must the design set be stored, but also the nearest few cases among maybe 100000 design set elements must be found to classify each applicant. However, advances in hardware and software mean that this can be done in mere seconds (see Henley and Hand (1996)).

4.2.9. *Time varying models*

In practical terms, application scoring models based on simplistic assumptions that applicants are good or bad risks are by far the most important. However, as we have already noted, they unrealistically oversimplify the situation. An applicant's propensity to default will vary over time as their circumstances vary. Bierman and Hausman (1970), Dirickx and Wakeman (1976) and Srinivasan and Kim (1987b) described a profit-based approach to determining whether or not to grant corporate credit on the basis of Bayesian updating of the default probability over time. Long (1976) also combined profitability and time evolution in a single model. Edelman (1992) studied the time evolution of delinquent accounts by using cluster analysis (and also, incidentally, showed why even defining bad risks is not a straightforward task). Cyert *et al.* (1962) modelled the time evolution of the distribution of amount due by time since repayment was due. Further work on this model is described by Corcoran (1978), van Kuelen *et al.* (1981), Frydman *et al.* (1985) and Mehta (1970).

4.3. *Which Method is Best?*

In general there is no overall 'best' method. What is best will depend on the details of the problem: on the data structure, the characteristics used, the extent to which it is possible to separate the classes by using those characteristics and the objective of the classification (overall misclassification rate, cost-weighted misclassification rate, bad risk rate among those accepted, some measure of profitability, etc.). If the classes are not well separated, then Pr(good risk|characteristic vector) is a rather flat function, so that the decision surface separating the classes will not be accurately estimated. In such circumstances, highly flexible methods such as neural networks and nearest neighbour methods are vulnerable to overfitting the design data and considerable smoothing must be used (e.g. a very large value for k, the number of nearest neighbours).

Classification accuracy, however measured, is only one aspect of performance. Others include the speed of classification, the speed with which a score-card can be revised and the ease of understanding of the classification method and why it has reached its conclusion. As far as the speed of classification goes, an instant decision is much more appealing to a potential borrower than is having to wait for several days.

Instant offers can substantially reduce the attrition rate. Robustness to population drift is attractive—and, when this fails, an ability to revise a score-card rapidly (and cheaply) is important. We have referred to the fact that nearest neighbour methods are effective in this regard. Classification methods which are easy to understand (such as regression, nearest neighbour and tree-based methods) are much more appealing, both to users and to clients, than are methods which are essentially black boxes (such as neural networks). They also permit more ready explanations of the sorts of reasons why the methods have reached their decisions.

Neural networks are well suited to situations where we have a poor understanding of the data structure. In fact, neural networks can be regarded as systems which combine automatic feature extraction with the classification process, i.e. they decide how to combine and transform the raw characteristics in the data, as well as yielding estimates of the parameters of the decision surface. This means that such methods can be used immediately, without a deep grasp of the problem. In general, however, if we do have a good understanding of the data and the problem, then methods which make use of this understanding might be expected to perform better. In credit scoring, where people have been constructing score-cards on similar data for several decades, there is solid understanding. This might go some way towards explaining why neural networks have not been adopted as regular production systems in this sector, despite the fact that banks have been experimenting with them for several years.

Because there is such a good understanding of the problem domain, it is very unlikely that new classification methodologies will lead to other than a tiny improvement in classification accuracy. In our experience, there is normally little to choose between the results of sensitive and sophisticated use of any of the methods. For example, Davis *et al.* (1992) described a comparison study of various techniques, including recursive partitioning and neural networks, and concluded that

> 'overall, they all perform at the same level of classification accuracy, but the neural network algorithms take much longer to train'.

In general, we believe that significant improvements are more likely to come from including new, more predictive, characteristics (or, of course, from changing the classification strategy—to, for example, using behavioural scoring in place of application scoring or to using risk-based pricing on loan offers).

We should consider what we mean by 'significant' in the last sentence. Henley and Hand (1996) developed an adaptive metric nearest neighbour method (with a parameter D describing the shape of the metric) for credit scoring. Among their results were those given in Table 2. The nearest neighbour methods are superior in this comparison. However, these figures are based on test set samples of about 5000

TABLE 2
Some results from Henley and Hand (1996)

Method	*Bad risk rate (%)*
k nearest neighbour (any D)	43.09
k nearest neighbour ($D = 0$)	43.25
Logistic regression	43.30
Linear regression	43.36
Decision graph or tree	43.77

with acceptance rates of 70%, so the percentages in Table 2 have denominators of about 3500. Using the last entry in Table 2 as a base-line, this means that the differences between the other methods and the last method, in terms of numbers of applicants are, in order from the top of Table 2, 24, 18, 16 and 14. These numbers are not very large, especially when put in the context of population drift and looseness of the good and bad risk class definitions. When one factors in the cost of changing the scoring system, and the likely future life of any system that one does install, one questions whether the differences are of any practical value.

5. REJECT INFERENCE

In practice, the design sample used to construct the classifier is rarely a random sample from the entire population. Typically, it is the set of people who were classified as good risks by an earlier score-card. Those in the 'reject' region were not granted credit—and hence were not followed up to determine their true risk status. All that is known about such people are the details given on their application forms (plus, perhaps, supplementary information about earlier loan repayment performance). This distortion of the distribution of applicants clearly has implications for the accuracy and general applicability of any new score-card that is constructed.

To allow for this, a widespread practice in the credit control industry is *reject inference*. This describes the practice of attempting to infer the likely true class of the rejected applicants and then using this information to yield a new score-card that is superior to one built on only those accepted for credit. Methods for reject inference are described in Hsia (1978) (the *augmentation* method), Reichert *et al.* (1983) and Joanes (1993). We can distinguish two cases. If the new score-card is based on a superset of the characteristics used in the original score-card then the true classes in the reject region are missing, but those in the accept region are not. In this case, the available data can be used to construct an accurate model, without taking into account the rejected cases, but only over the 'accept' regions of the space, as defined by the original classifier. Extrapolation over the former reject region is then needed. However, if the new score-card does not include all the characteristics used in the original score-card then the true classes among those which have been rejected are *non-ignorably* missing (in the terminology introduced by Little and Rubin (1987)). In this case, the observed distribution of good or bad risks is not representative of the true distribution. We might try to overcome this by including a model for the selection process in estimating the parameters of the new classification rule (see, for example, Copas and Li (1997)). However, since the new rule does not include all the characteristics used for the original rule it is unlikely to outperform the original rule. Hand and Henley (1993, 1994) have explored reject inference in detail. They concluded that reject inference cannot work unless additional assumptions were made, such as assuming particular forms for the distributions of the good and bad risks. A suggestion that this was the case was made by Reichert *et al.* (1983) who concluded that

> 'the inclusion of the group of rejected applicants appears to have little information that is useful in classifying marginal credit risks'.

Improved classification rules could be produced if information was available in the reject region—if some applicants who would normally be rejected were accepted.

This would be a commercially sensible thing to do if the loss due to the increased number of delinquent accounts was compensated for by the increased accuracy in classification. Rosenberg and Gleit (1994) refer to one company that initially grants everyone a small amount of credit and we know of an organization which accepts a sample from the reject region. The practice seems very rare, however. A related practice, which is increasingly common, is to obtain information on rejected applicants from other credit suppliers who did grant them credit.

6. LEGAL ASPECTS

Legislation prevents the use of certain characteristics, such as sex or race, in the credit granting decision. One viewpoint is that the aim of this is to ensure that no irrational prejudices influence the decision, so that the classification is solely on merit with respect to the objective of the classification (credit risk). However, if classification is to be based solely on merit, one might argue that it would be appropriate to seek the best risk predictor that one could find, which would naturally include all characteristics thought to influence risk, perhaps including those currently prohibited. To do otherwise, one could argue, would mean that some groups were inevitably being unjustly penalized, in that they were being assigned default risk probabilities that were greater than their true probabilities (ignoring issues of accuracy of estimation).

Alternatively, one might argue that the objective is to eliminate both the direct influence and the indirect influence of subgroup membership from the credit scoring decision. The current practice of outlawing consideration of subgroup membership addresses only the first of these—and permits variables which can act as (partial) proxies for the excluded variables. Direct and indirect influence can both be eliminated if separate score-cards are constructed for each subgroup, with the thresholds being chosen so that the same proportions are accepted within each subgroup.

Constraints on the information which may or may not be used in constructing a credit classification rule are one class of legislative restrictions. The Consumer Credit Act (1974) also requires credit reference agencies to divulge information to individuals on request, and to remove or correct it if it is incorrect.

7. OTHER ISSUES

This paper has focused primarily on the classification of applicants into good or bad risk classes on the basis of their initial application characteristics and has only touched on other areas. However, there are many other areas of credit scoring and credit control which also present interesting statistical challenges, such as the following.

(a) Loan servicing and review functions: for example, Blackwell and Sykes (1992) have described the use of behavioural scoring to determine credit limits. There are also questions such as when to approach customers with an invitation to top up their loans.
(b) By risk-based pricing, in which the interest rate charged varies according to the estimated risk, one can, in principle at least, never turn down a loan application. Techniques such as this depend heavily on the computer.

(c) Fraud is an area of increasing interest to credit grantors. Leonard (1993c) described an expert system aimed at detecting the fraudulent use of credit cards and Henley (1995) described an attempt to build a fraud score-card by using linear regression analysis.

(d) We have referred, at various places in the paper, to profitability scoring. There is much scope for very effective modelling in this area, as well as for the development of flexible financial tools to increase profits.

(e) Questions also arise on when and how to act on a delinquent loan. First, is it worthwhile to pursue a delinquent loan — will the expected pay-off exceed the cost? Secondly, what action should be taken (a reminder letter or legal action?) and should it be taken at the first hint of trouble or will this needlessly antagonize customers who will pay? Score-cards can be (and generally have been) used for all such problems.

(f) A rather different application is the use of statistical methods to decide whom to invite to apply for a loan in the first place — essentially a marketing exercise. Given that the positive response rates for marketing exercises can be as low as 1% or 2%, the potential for improvement is vast. Score-cards and other predictive statistical methods have been used in this application, as have other rather different classes of techniques, such as cluster analysis for market segmentation (e.g. Lundy (1992)).

8. CONCLUSION

The major part of the statistical work in credit scoring and credit control to date has focused on the conceptually relatively straightforward aspect of constructing improved discriminating rules. However, it seems likely that the greatest advances in the future will occur in the development of more complex and sophisticated models, addressing issues such as those outlined in Section 7. The problems are basically statistical and certainly present challenging opportunities for statisticians.

ACKNOWLEDGEMENTS

We would like to express our appreciation for their constructive comments on earlier versions of this paper to Ross Gayler, Gerard Scallan, the referees and the journal editors. Their comments have led to substantial clarification and improvement. The work of the second author was supported by a research studentship grant from Littlewoods Plc.

REFERENCES

Apilado, V. P., Warner, D. C. and Dauten, J. J. (1974) Evaluative techniques in consumer finance. *J. Finan. Quant. Anal.*, Mar., 275–283.

Bierman, Jr, H. and Hausman, W. H. (1970) The credit granting decision. *Mangmnt Sci.*, **16**, 519–532.

Blackwell, M. and Sykes, C. (1992) The assignment of credit limits with a behaviour-scoring system. *IMA J. Math. Appl. Bus. Indstry*, **4**, 73–80.

Boyle, M., Crook, J. N., Hamilton, R. and Thomas, L. C. (1992) Methods for credit scoring applied to slow payers. In *Credit Scoring and Credit Control* (eds L. C. Thomas, J. N. Crook and D. B. Edelman), pp. 75–90. Oxford: Clarendon.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Belmont: Wadsworth.

Capon, N. (1982) Credit scoring systems: a critical analysis. *J. Marktng*, **46**, 82–91.

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

Chandler, G. G. and Coffman, J. Y. (1979) A comparative analysis of empirical versus judgemental credit evaluation. *J. Retail Bank.*, **1**, no. 2, 15–26.

Chatterjee, S. and Barcun, S. (1970) A nonparametric approach to credit screening. *J. Am. Statist. Ass.*, **65**, 150–154.

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

Copas, J. B. and Li, H. G. (1997) Inference for non-random samples (with discussion). *J. R. Statist. Soc.* B, **59**, 55–95.

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

Crook, J. N., Hamilton, R. and Thomas, L. C. (1992) A comparison of discriminators under alternative definitions of credit default. In *Credit Scoring and Credit Control* (eds L. C. Thomas, J. N. Crook and D. B. Edelman), pp. 217–245. Oxford: Clarendon.

Cyert, R. M., Davidson, H. J. and Thompson, G. L. (1962) Estimation of the allowance for doubtful accounts by Markov chains. *Mangmnt Sci.*, Aug., 287–303.

Davis, D. B. (1987) Artificial intelligence goes to work. *High Technol*, Apr., 16–17.

Davis, R. H., Edelman, D. B. and Gammerman, A. J. (1992) Machine-learning algorithms for credit-card applications. *IMA J. Math. Appl. Bus. Indstry*, **4**, 43–51.

Dirickx, Y. M. I. and Wakeman, L. (1976) An extension of the Bierman–Hausman model for credit granting. *Mangmnt Sci.*, **22**, 1229–1237.

Duffy, W. (1977) The credit scoring movement. *Credit*, Sept., 28–30.

Durand, D. (1941) *Risk Elements in Consumer Instalment Financing*. New York: National Bureau of Economic Research.

Edelman, D. B. (1992) An application of cluster analysis in credit control. *IMA J. Math. Appl. Bus. Indstry*, **4**, 81–87.

Edmister, R. O. and Schlarbaum, G. G. (1974) Credit policy in lending institutions. *J. Finan. Quant. Anal.*, June, 335–356.

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

——(1978) Problems in applying discriminant analysis in credit scoring models. *J. Bank. Finan.*, **2**, 205–219.

Fitzpatrick, D. B. (1976) An analysis of bank credit card profit. *J. Bank Res.*, **7**, 199–205.

Frydman, H., Kallberg, J. G. and Kao, D.-L. (1985) Testing the adequacy of Markov chains and mover-stayer models as representations of credit behaviour. *Ops Res.*, **33**, 1203–1214.

Gifi, A. (1990) *Nonlinear Multivariate Analysis*. Chichester: Wiley.

Grablowsky, B. J. and Talley, W. K. (1981) Probit and discriminant functions for classifying credit applicants: a comparison. *J. Econ. Bus.*, **33**, 254–261.

Greer, C. C. (1967) The optimal credit acceptance scheme. *J. Finan. Quant. Anal.*, **3**, 399–415.

Hand, D. J. (1981) *Discrimination and Classification*. Chichester: Wiley.

——(1985) *Artificial Intelligence and Psychiatry*. Cambridge: Cambridge University Press.

——(1986) New instruments for identifying good and bad credit risks: a feasibility study. *Report*. Trustee Savings Bank, London.

——(1994) Assessing classification rules. *J. Appl. Statist.*, **21**, 3–16.

Hand, D. J. and Henley, W. E. (1993) Can reject inference ever work? *IMA J. Math. Appl. Bus. Indstry*, **5**, 45–55.

——(1994) Inference about rejected cases in discriminant analysis. In *New Approaches in Classification and Data Analysis* (eds E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and B. Burtschy), pp. 292–299. New York: Springer.

Hand, D. J., McConway, M. J. and Stanghellini, E. (1996a) Graphical models of applicants for credit. *IMA J. Math. Appl. Bus. Indstry*, to be published.

Hand, D. J., Oliver, J. J. and Lunn, A. D. (1996b) Discriminant analysis when the classes arise from a continuum.

Henley, W. E. (1995) Statistical aspects of credit scoring. *PhD Thesis*. The Open University, Milton Keynes.

Henley, W. E. and Hand, D. J. (1996) A k-nearest-neighbour classifier for assessing consumer credit risk. *Statistician*, **45**, 77–95.

Hsia, D. C. (1978) Credit scoring and the equal credit opportunity act. *Hast. Law J.*, **30**, 371–448.

Joanes, D. N. (1993) Reject inference applied to logistic regression for credit scoring. *IMA J. Math. Appl. Bus. Indstry*, **5**, 35–43.

Kolesar, P. and Showers, J. L. (1985) A robust credit screening model using categorical data. *Mangmnt Sci.*, **31**, 123–133.

van Kuelen, J. A. M., Spronk, J. and Corcoran, A. W. (1981) On the Cyert–Davidson–Thompson doubtful accounts model. *Mangmnt Sci.*, **27**, 108–112.

Lachenbruch, P. A. (1975) *Discriminant Analysis*. New York: Hafner.

Lane, S. (1972) Submarginal credit risk classification. *J. Finan. Quant. Anal.*, **7**, 1379–1385.

Leonard, K. J. (1993a) Empirical Bayes analysis of the commercial loan evaluation process. *Statist. Probab. Lett.*, **18**, 289–296.

——(1993b) Detecting credit card fraud using expert systems. *Comput. Indstrl Engng*, **25**, 103–106.

——(1993c) A fraud-alert model for credit cards during the authorization process. *IMA J. Math. Appl. Bus. Indstry*, **5**, 57–62.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

Long, M. S. (1976) Credit screening system selection. *J. Finan. Quant. Anal.*, June, 313–328.

Lucas, A. (1992) Updating scorecards: removing the mystique. In *Credit Scoring and Credit Control* (eds L. C. Thomas, J. N. Crook and D. B. Edelman), pp. 180–197. Oxford: Clarendon.

Lundy, M. (1992) Cluster analysis in credit scoring. In *Credit Scoring and Credit Control* (eds L. C. Thomas, J. N. Crook and D. B. Edelman), pp. 91–107. Oxford: Clarendon.

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

McGuire, C. H. (1985) Medical problem solving: a critique of the literature. *J. Med. Educ.*, **60**, 587–595.

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

——(1970) Optimal credit policy selection: a dynamic approach. *J. Finan. Quant. Anal.*, Dec., 421–444.

Moses, D. and Liao, S. S. (1987) On developing models for failure prediction. *J. Commrcl Bank Lend.*, **69**, 27–38.

Myers, J. H. and Forgy, E. W. (1963) The development of numerical credit evaluation systems. *J. Am. Statist. Ass.*, **58**, 799–806.

Oliver, R. M. (1992) The economic value of score-splitting accept-reject policies. *IMA J. Math. Appl. Bus. Indstry*, **4**, 35–41.

Orgler, Y. E. (1970) A credit scoring model for commercial loans. *J. Money Credit Bank.*, Nov., 435–445.

——(1971) Evaluation of bank consumer loans with credit scoring models. *J. Bank Res.*, spring, 31–37.

Reichert, A. K., Cho, C.-C. and Wagner, G. M. (1983) An examination of the conceptual issues involved in developing credit-scoring models. *J. Bus. Econ. Statist.*, **1**, 101–114.

Ripley, B. D. (1994) Neural networks and related methods for classification (with discussion). *J. R. Statist. Soc.* B, **56**, 409–456.

Rosenberg, E. and Gleit, A. (1994) Quantitative methods in credit management: a survey. *Ops Res.*, **42**, 589–613.

Safavian, S. R. and Landgrebe, D. (1991) A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cyb.*, **21**, 660–674.

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

Srinivasan, V. and Kim, Y. H. (1987a) Credit granting: a comparative analysis of classification procedures. *J. Finan.*, **42**, 665–683.

——(1987b) The Bierman–Hausman credit granting model: a note. *Mangmnt Sci.*, **33**, 1361–1362.

Wiginton, J. C. (1980) A note on the comparison of logit and discriminant models of consumer credit behaviour. *J. Finan. Quant. Anal.*, **15**, 757–770.

Wilkie, A. D. (1992) Measures for comparing scoring systems. In *Credit Scoring and Credit Control* (eds L. C. Thomas, J. N. Crook and D. B. Edelman), pp. 123–138. Oxford: Clarendon.

Zocco, D. P. (1985) A framework for expert systems in bank loan management. *J. Commrcl Bank Lend.*, **67**, 47–54.

Zweig, M. H. and Campbell, G. (1993) Receiver-operating characteristic (ROC) plots. *Clin. Chem.*, **29**, 561–577.