Available online at www.sciencedirect.com
ScienceDirect

European Journal of Operational Research 183 (2007) 1447–1465

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/ejor

Recent developments in consumer credit risk assessment

Jonathan N. Crook [a,*], David B. Edelman [b], Lyn C. Thomas [c]

[a] *Credit Research Centre, Management School and Economics, 50 George Square, University of Edinburgh, Edinburgh EH8 9JY, United Kingdom*
[b] *Caledonia Credit Consultancy, 42 Milverton Road, Glasgow G46 7LP, United Kingdom*
[c] *School of Management, Highfield, Southampton SO17 1BJ, United Kingdom*

Received 1 July 2006; accepted 1 September 2006
Available online 15 February 2007

Abstract

Consumer credit risk assessment involves the use of risk assessment tools to manage a borrower's account from the time of pre-screening a potential application through to the management of the account during its life and possible write-off. The riskiness of lending to a credit applicant is usually estimated using a logistic regression model though researchers have considered many other types of classifier and whilst preliminary evidence suggest support vector machines seem to be the most accurate, data quality issues may prevent these laboratory based results from being achieved in practice. The training of a classifier on a sample of accepted applicants rather than on a sample representative of the applicant population seems not to result in bias though it does result in difficulties in setting the cut off. Profit scoring is a promising line of research and the Basel 2 accord has had profound implications for the way in which credit applicants are assessed and bank policies adopted.

Keywords: Finance; OR in banking; Risk analysis

1. Introduction

Between 1970 and 2005 the volume of consumer credit outstanding in the US increased by 231% and the volume of bank loans secured on real estate increased by 705%.[1] Of the $3617.0 billion of outstanding commercial bank loans secured on real estate and of consumer loans in the personal sector in December 2005, the former made up 80%.[2] The growth in debt outstanding in the UK has also been dramatic. Between 1987 and 2005 the volume of consumer credit (that is excluding mortgage debt) outstanding increased by 182% and the growth in credit card debt was 416%.[3] Mortgage debt

[*] Corresponding author. Tel.: +44 131 6503820; fax: +44 131 6683053.

E-mail addresses: j.crook@ed.ac.uk (J.N. Crook), davide24@ntlworld.com (D.B. Edelman), l.thomas@soton.ac.uk (L.C. Thomas).

[1] Data from the Federal Reserve Board (series bcablci_ba.m and bcablcr_ba.m deflated), H8, *Assets and Liabilities of Commercial Banks in the United States.*

[2] Data from the Federal Reserve Board (series bcablci_ba.m and bcablcr_ba.m deflated), H8, *Assets and Liabilities of Commercial Banks in the United States.*

[3] Data from ONS Online.

0377-2217/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2006.09.100

outstanding increased by 125%.[4] Within Europe growth rates have varied. For example between 2001 and 2004 nominal loans to households and the non-profit sector increased by 36% in Italy, but in Germany they increased by only 2.3% with the Netherlands and France at 28% and 22% respectively.[5]

These generally large increases in lending and associated applications for loans have been underpinned by one of the most successful applications of statistics and operations research: credit scoring. Credit scoring is the assessment of the risk associated with lending to an organization or an individual. Every month almost every adult in the US and the UK is scored several times to enable a lender to decide whether to mail information about new loan products, to evaluate whether a credit card company should increase one's credit limit, and so on. Whilst the extension of credit goes back to Babylonian times (Lewis, 1992) the history of credit scoring begins in 1941 with the publication by Durand (1941) of a study that distinguished between good and bad loans made by 37 firms. Since then the already established techniques of statistical discrimination have been developed and an enormous number of new classificatory algorithms have been researched and tested. Virtually all major banks use credit scoring with specialised consultancies providing credit scoring services and offering powerful software to score applicants, monitor their performance and manage their accounts. In this review we firstly explain the basic ideas of credit scoring and then discuss a selection of very exciting current research topics. A number of previous surveys have been published albeit with different emphasises. Rosenberg and Gleit (1994) review different types of classifiers, (Hand and Henley, 1997) cover earlier results concerning the performance of classifiers whilst (Thomas, 2000) additionally reviews behavioural and profit scoring. Thomas et al. (2002) discuss research questions in credit scoring especially those concerning Basel 2.

2. Basic ideas of credit scoring

Consumer credit risk assessment involves the use of risk assessment tools to manage a borrower's account from the time of direct mailing of marketing material about a consumer loan through to the management of the borrower's account during its lifetime. The same techniques are used in building these tools even though they involve different information and are applied to different decisions. Application scoring helps a lender discriminate between those applicants whom the lender is confident will repay a loan or card or manage their current account properly and those applicants about whom the lender is insufficiently confident. The lender uses a rule to distinguish between these two subgroups which make up the population of applicants. Usually the lender has a sample of borrowers who applied, were made an offer of a loan, who accepted the offer and whose subsequent repayment performance has been observed. Information is available on many sociodemographic characteristics (such as income and years at address) of each borrower at the time of application from his/her application form and typically on the repayment performance of each borrower on other loans and of individuals who live in the same neighbourhood. We will denote each characteristic as x_i which may take on one of several values (called "attributes") x_{ij} for case i.

The most common method of deriving a classification rule is logistic regression where, using maximum likelihood, the analyst estimates the parameters in the equation:

$$\mathrm{Log}\left(\frac{p_{gi}}{1-p_{gi}}\right) = \beta_0 + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x_i}, \tag{1}$$

where p_{gi} is the probability that case i is a good.

This implies that the probability of case i being a good is

$$p_{gi} = \frac{\mathrm{e}^{\boldsymbol{\beta}^{\mathbf{T}}\mathbf{x_i}}}{1+\mathrm{e}^{\boldsymbol{\beta}^{\mathbf{T}}\mathbf{x_i}}}.$$

Traditionally the values of the characteristics for each individual were used to predict a value for p_{gi} which is compared with a critical or cut off value and a decision made. Since a binary decision is required the model is required to provide no more than a ranking. p_{gi} may be used in other ways. For example in the case of a card product, p_{gi} will also determine a multiple of salary for the credit limit. For a current account, p_{gi} will determine the number of cheques in a cheque book and the type of card that is issued. For a loan product p_{gi} will often determine the interest rate to be charged.

In practice the definition of a good varies enormously but is typically taken as a borrower who

[4] Data from the Council of Mortgage Lenders.

[5] Source: OECDStatistics, Financial Balance Sheets – consolidated dataset 710.

defaulted within a given time period, usually within one year. Default may be taken as missed three scheduled payments in the period or missed three successive payments within the period.

In practice the characteristics in Eq. (1) may be transformed. Typically each variable, whether measured at nominal, ordinal or ratio level is divided into ranges according to similarity of the proportion of goods at each value. Inferential tests are available to test the null hypothesis that any one aggregation of values into a range separates the goods from the bads more clearly than another aggregation. This process is known as "coarse classification" and the available tests include the chi-square test, the information statistic and the Somer's D concordance statistic. A transformation of the proportion that is good is performed for the cases in each band, a typical example being to take the log of the proportion that are good divided by the proportion that are bad. The transformed value, called the weight of evidence, is used in place of the raw values for all cases in the range. This has the advantage of preserving degrees of freedom in a small dataset. However most datasets are sufficiently large that this method confers few benefits over using either dummy variables to represent each range (nominal and continuous variables) or dummy variables. Coarse classification also has the advantage of allowing a missing value (i.e. no response to a question) to be incorporated in the analysis of the responses to that question.

The model is parameterised on a training sample and tested on a holdout sample to avoid over-parameterisation whereby the estimated model fits the nuances in the training sample which are not repeated in the population. Validation of an application scoring model involves assessing how well it separates the distributions of observed goods and bads over the characteristics: its discriminatory power. Fig. 1 shows two such observed distributions produced from a scoring model like Eq. (1). S_c is a cut-off score used to partition the score range into those to be labelled goods (i.e. to be accepted), this group to be labelled A_G, and those to be labelled bads (and so rejected), labelled A_B. If the distributions did not overlap the separation and so the discrimination would be perfect. If they do overlap there are four subsets of cases. Bads classified as bad, bads classified as good, goods classified as good and goods classified as bad. The proportion of bads classified as good is the area under the B curve to the right of S_c, labelled B^W. The proportion of goods classified as bad is the area under the G curve to the left of S_c, labelled G^W. One measure used in practice is to estimate the sum of these proportions for a given cut-off. A two by two matrix of the four possibilities is constructed, called a confusion matrix, an example of which is shown in Table 1. In this table G^W is $n_{BG}/(n_{GG}+n_{BG})$ and B^W is $n_{GB}/(n_{GB}+n_{BB})$. The cut off may be set in many ways for example such that the error rate is minimised or equal to the observed error rate in the training sample or in the holdout sample. Notice that the appropriate test sample is the population of all cases whereas in practice this is not available and we must make do with a sample. Now the expected value of the error rate estimated from all holdout samples is an unbiased estimate of the error rate in the population, if the sample is selected randomly from all possible applicants. However since, in practice, the sample typically consists only of accepted applicants, the expected error rate from all holdout samples may be a biased estimate of the error rate in the population of all applicants (see reject inference below).

A weakness of the error rate is that the predictive accuracy of the estimated model depends on the cut-off value chosen. On the other hand, the statistics associated with the receiver operating curve (ROC) give a summary measure of the discrimination of

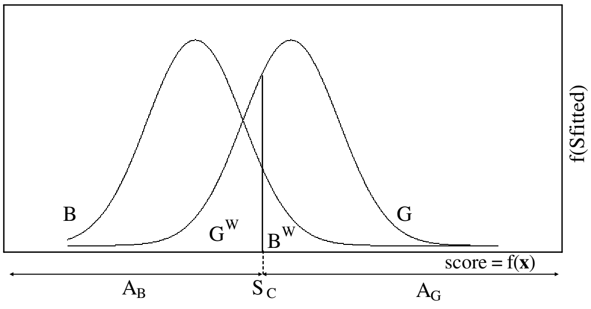

Fig. 1.

Table 1
A confusion matrix

Predicted class	Observed class		
	Good	Bad	Row total
Good	n_{GG}	n_{GB}	$n_{GG}+n_{GB}$
Bad	n_{BG}	n_{BB}	$n_{BG}+n_{BB}$
Column total	$n_{GG}+n_{BG}$	$n_{GB}+n_{BB}$	$n_{GG}+n_{BG}+$ $n_{GB}+n_{BB}$

the scorecard since all possible cut-off scores are considered. The ROC curve is a plot of the proportion of bads classified as bad against the proportion of goods classified as bad at all values of S_c. The former is called "sensitivity" and the latter is 1 – "specificity" since specificity is the proportion of goods predicted to be good. We can see these values in Table 1 where sensitivity is $n_{\mathrm{BB}}/(n_{\mathrm{GB}}+n_{\mathrm{BB}})$ and 1-specificity is $n_{\mathrm{BG}}/(n_{\mathrm{GG}}+n_{\mathrm{BG}})$. If the value of S_c is increased from its lowest possible value the value of B^W increases faster than the value of G^W decreases, eventually the reverse happens. This gives the shape of the ROC curve shown in Fig. 2. If the distributions are totally separated, as S_c increases, all of the bads will be correctly classified before any of the goods are misclassified and the ROC curve will lie over the vertical axis until point A and then along the upper horizontal axis. If there is no separation at all and the distributions are identical, the proportion of bads classified as bad will equal the portion of goods classified as bad and the ROC curve will lie over the diagonal straight line. One measure of the separation yielded by a scoring model is therefore the proportion of the area below the ROC curve which is above the diagonal line. The Gini coefficient is double this value. The Gini coefficient can be compared between scoring models to indicate which gives better separation. If one is interested in predictive performance over all cut-off values the Gini coefficient is more informative than looking at ROC plots since the latter may cross. A useful property is that the area under the ROC curve down to the right hand corner of Fig. 2 equals the Mann–Whitney statistic. Assuming this is normally distributed one can place confidence intervals on this area. But the Gini may be very misleading when we are interested only in a small range of cut-off values around where the predicted bad rate is acceptable because the Gini is the integral over all cut-offs. When we are interested in predictive performance only in a narrow range of cut-offs, the ROC plots would give more information.

Other important measures of separation are available. One is the Kolmogorov–Smirnov statistic (KS). From Fig. 1 we can plot the cumulative distributions of scores for goods and bads from the scoring model, which are shown in Fig. 3, where $F(s|G)$ denotes the probability that a good has a score less than s and $F(s|B)$ denotes the probability that a Bad has a score less than s. The KS is the maximum difference between $F(s|G)$ and $F(s|B)$ at any score. Intuitively the KS says: across all scores what is the maximum difference between the probability that a case is a good and is rejected and the probability one is a bad and is rejected. However, like the Gini, the KS may also give the maximum separation at cut-offs that the lender is, for operational reasons, not interested in. Many other measures are available; see (Thomas et al., 2002; Tasche, 2005).

Hand (2005) has cast doubt on the value of any measure of scorecard performance which uses the distribution of scores. Since the aim of, say, an application model is simply to divide applicants into those to be accepted and the rest, it is only the numbers of cases above and below the cut-off that is relevant. The distance between the score and the cut-off is not. The proportion of accepted bad cases uses just this information and so is to be preferred over the area under the ROC curve, Gini or KS measure. If on the other hand, we are interested in dividing a sample into two groups where we can follow the subsequent ex post performance of both groups, then we can find the proportion of bad cases that turned out to be good and the proportion of good cases that turned out to be bad. Given the costs of each type of misclassification the total such

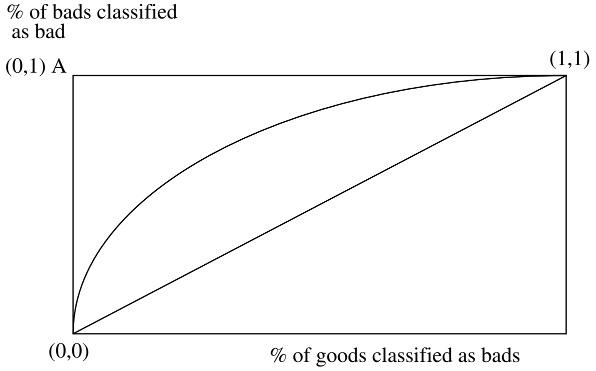

Fig. 2. The receiver operating curve.

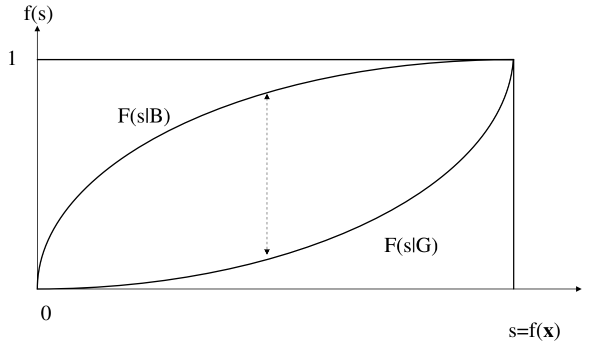

Fig. 3. The Kolmogrov–Smirnov statistic.

costs from a scorecard can be estimated. Again only the sign of the classification and the costs of misclassification are used, the distribution of scores is not.

Once an applicant has been accepted and has been through several repayment cycles a lender may wish to decide if he/she is likely to default in the next time period. Unlike application scoring the lender now has information on the borrower's repayment performance and usage of the credit product over time period t and may use this in addition to the application characteristic values to predict the chance of default over time period $t+1$. This may be used to decide whether to increase a credit limit, grant a new loan or even post a mail shot. A weakness of this approach is that it assumes that a model estimated from a set of policies towards the borrower in the past (size of loan or limit) is just as applicable when the policy is changed (on the basis of the behavioural score – see below). This may well be false.

Another approach is to model the behaviour of the borrower as a Markov chain (MC). Each transition probability, $p_t(i,j)$ is the probability that an account moves from state i to state j one month later. The repayment states may be "no credit", "has credit and is up to date", "has credit and is one payment behind" ... "default". Typically "written off" and "loan paid off" are absorption states. Alternatively MCs may be more complex. Many banks may use a compound definition for example how many payments are missed, whether the account is active, how much money is owed and so on. A good example, a model developed by Bank One, is given by (Trench et al., 2003). Notice that there are typically many structural zeros: transitions which are not possible between two successive months. For example an account cannot move from being up to date to being more than one period behind.

Those transition matrices that have been published typically show that, apart from those who have missed one or more payments, at least 90% of credit card accounts (Till and Hand, 2001) or personal loan accounts (Ho et al., 2004) remain up to date, that is move from up to date in period t to up to date in period $t+1$. Till and Hand (2001) found that for credit cards, apart from those who missed 1 or 2 payments in t, the majority of those who missed 3 or more payments in t missed a further payment in $t+1$ and as we consider more missed payments, this proportion of cases goes up until five missed payments when the proportion levels off. They also found that the proportion of cases in a state that move directly to paid up to date decreases up until five missed payments when it increases again. This last result may be due to people taking out a consolidating loan to pay off previous debts.[6]

If the transition probability matrix, $\boldsymbol{P_t}$, is known for each t one can predict the distribution of the states of accounts that a customer has n periods into the future. A kth order non-stationary MC may be written as

$$\begin{aligned}P\{X_{t+k} &= j_{t+k} | X_{t+k-1} = j_{t+k-1}, X_{t+k-2} = j_{t+k-2}, \\ &\qquad \ldots\ldots\ldots, X_1 = j_1\} \\ &= p_t(j_t, \ldots\ldots, j_{t+k-1}, j_{t+k}) \\ &= p_t(\{j_t, \ldots\ldots, j_{t+k-1}\}, \{j_{t+1}, \ldots, j_{t+k}\}),\end{aligned} \tag{2}$$

where X_t takes on values representing the different possible repayment states.

A MC is stationary if the elements within the transition matrix do not vary with time. A kth order MC is one where the probability that a customer is in a state j in period $t+k$ depends on the states that the customer has visited in the previous k periods. If an MC is non-stationary (but first order) the probability that a customer moves from state i to state j over k periods is the product of the transition matrices which each contain the transition probabilities between two states for every two adjacent periods. One may represent a kth order MC by redefining each "state" to consist of every possible combination of values of $X_t, X_{t+1}, \ldots, X_{t+k}$ originally defined states. Then the transition matrix of probabilities over t to $t+k$ periods is the product of the matrix containing transition probabilities over the $(t, \ldots, t+k-1)$ periods and $(t+1, \ldots, t+k\text{th})$ period. Anderson and Goodman (1957) give tests for stationarity and order. Evidence suggests that MCs describing repayment states are not first order and may not even be second order (Till and Hand, 2001). Ho et al. (2004) find that a MC describing states for a bank are not first, second or third order, where each state is defined as a combination of how delinquent and how active the account is and the outstanding balance owed.

[6] Or to the fact that some systems actually have the original debt cleared as the case is passed to an external debt collection agency or written off.

Results by Till and Hand (2001) also suggest MCs for the delinquency states for credit products are not stationary. Trench et al. (2003) consider redefining the states by concatenation of previous historical states to induce stationarity, but opt instead for segmenting the transition matrix for the whole sample into a matrix for each customer segment. The resulting model, when combined with an optimisation routine to maximise NPV for each exiting account by choice of credit line and APR, has considerable success.

Often MC models are used for predicting the long term distributions of accounts and so enabling policies to be followed to change these distributions. Early applications of MCs (Cyert and Thompson, 1968) to credit risk involved segmenting borrowers into risk groups with each group having its own transition matrix. More recent applications have involved Mover–Stayer Markov chains which are chains where the population of accounts is segmented into those which "stay" in the same state and those which move (Frydman, 1984). So the j period transition matrix consists of the probability that an account is a stayer times the transition matrix given that it is a stayer, plus the probability it is a mover times the transition matrix given that it is a mover. So (if we assume stationarity):

$$\mathbf{P}(\mathbf{0},\mathbf{j}) = \mathbf{SI} + (\mathbf{I} - \mathbf{S})\mathbf{M}^{j}, \tag{3}$$

where $\mathbf{S}$ = row vector consisting of the diagonal of the matrix of S_i values where S_i is the proportion of stayers in state i. $\mathbf{M}$ is the transition matrix for movers and $\mathbf{I}$ is the identity matrix. Notice that the stayers could be in any state of the transition matrix, and not just the up to date state. Studies show that between 44% and 72% of accounts never leave the same state (the bulk of which are in the up to date state), which supports the applicability of a stayer segment in credit portfolios. Recent research has decomposed the remaining accounts into several segments rather than just "movers". Thus Ho and Thomas were able to statistically separate the movers into three groups according to the number of changes of state made: those that move modestly "twitchers", those that move more substantially "shakers" and those that move considerably "movers".

3. Classification algorithms

We now begin our discussion of a selection of areas where recent research has been particularly active. There are many other areas of rapid research in this area and we do not mean to imply that the areas we have chosen are necessarily the most beneficial areas.

3.1. Statistical methods

Whilst logistic regression (LR) is a commonly used algorithm it was not the first to be used and the development of newer methods has been a long standing research agenda which is likely to continue. One of the earliest techniques was discriminant analysis. Using our earlier notation we wish to develop a rule to partition a sample of applicants, A, into goods A_G and bads A_B which maximises the separation between the two groups in some sense. Fisher (1936) argued that a natural solution is to find the direction, $\mathbf{w}$ that maximises the difference between the sample mean values of $\mathbf{x}_G$ and $\mathbf{x}_B$. If we define $y_G = \mathbf{w}^T\mathbf{x}_G$ and $y_B = \mathbf{w}^T\mathbf{x}_B$ then we wish to maximise $E(y_G - y_B)$ and so to maximise $\mathbf{w}^T(\overline{\mathbf{x}_G} - \overline{\mathbf{x}_B})$. In terms of Fig. 1 we wish to find the $\mathbf{w}$ matrix that maximises the difference between the peaks of the B and G distributions in the horizontal direction. But although the sample means may be widely apart, the distributions of $\mathbf{w}^T\mathbf{x}$ values may overlap considerably. The misclassifications in Fig. 1 represented by the G^W and B^W areas would be relatively large, but for a given difference in means would be lower if the distributions had smaller variances. If, for simplicity, we assume the two groups have the same covariance matrices, denoted as $\mathbf{S}$ in the samples, the within-group variance is $\mathbf{w}^T\mathbf{S}\mathbf{w}$ and dividing $\mathbf{w}^T(\overline{\mathbf{x}_G} - \overline{\mathbf{x}_B})$ by this makes the separation between the two group means smaller for larger variances. In fact the distance measure that is used is

$$\mathbf{M}(\mathbf{w}) = \frac{[\mathbf{w}^T(\overline{\mathbf{x}_G} - \overline{\mathbf{x}_B})]^2}{\mathbf{w}^T\mathbf{S}\mathbf{w}}, \tag{4}$$

(where the numerator is squared for mathematical convenience). Differentiating equation (4) with respect to $\mathbf{w}$ gives estimators for $\mathbf{w}$ where $\mathbf{w} \propto \mathbf{S}^{-1}$ $(\overline{\mathbf{x}_G} - \overline{\mathbf{x}_B})$. Since we just require a ranking of cases to compare with a cut-off value this is sufficient to classify each case. $\mathbf{S}$ is taken as an estimate of the population value $\boldsymbol{\Omega}$ and $\overline{\mathbf{x}_G}$ and $\overline{\mathbf{x}_B}$ are taken as estimates of the corresponding population values, $\boldsymbol{\mu}_G$ and $\boldsymbol{\mu}_B$.

Notice that we have made no assumptions about the distributions of $\mathbf{x}_G$ or $\mathbf{x}_B$. This is particularly important for credit scoring because such assumptions rarely hold even when weights of evidence values are used.

One can in fact deduce the same estimator in other ways. For example we could partition the sample to minimise the costs of doing this erroneously. That is, in Fig. 1 minimise the sum of areas G^W and B^W, each weighted by the cost of misclassification of each case. Assuming that the values of the characteristics, $\mathbf{x}$ are multivariate normal and they have a common covariance matrix we have

$$A_{\mathrm{G}} : y \leqslant \mathbf{x}^{\mathbf{T}}\mathbf{w}, \tag{5}$$

where $\mathbf{w} = \Omega^{-1}(\boldsymbol{\mu}_{\mathrm{G}} - \boldsymbol{\mu}_{\mathrm{B}})$ and $y = 0.5(\boldsymbol{\mu}_{\mathrm{G}}^{\mathbf{T}} \cdot \Omega^{-1} \cdot \boldsymbol{\mu}_{\mathrm{G}} - \boldsymbol{\mu}_{\mathrm{B}}^{\mathbf{T}}\Omega^{-1}\boldsymbol{\mu}_{\mathrm{B}}) + \log\left(\frac{Dp_{\mathrm{B}}}{Lp_{\mathrm{G}}}\right)$ where D is the expected loss when a bad is accepted and L is the expected foregone profit when a good is rejected. For estimation the population means $\boldsymbol{\mu}_{\mathrm{G}}$ and $\boldsymbol{\mu}_{\mathrm{B}}$ and the covariance matrix are estimated by their sample values.

Another common rule is a simple linear model estimated by Ordinary Least Squares where the dependent variable takes on the values of 1 or 0 according to whether the applicant has defaulted or not. Unfortunately this has the disadvantage that predicted probabilities can lie outside the range $[0, 1]$ which is a characteristic not shared with logistic regression. This may not matter if one only requires a ranking of probabilities of default, but may be problematic if an estimated probability of default (PD) is required for capital adequacy purposes.

3.2. Mathematical programming

More recently non-statistical methods have been developed. Mathematical programming derived classifiers have found favour in several consultancies. Here one wishes to use MP to estimate the weights in a linear classifier so that the weighted sum of the application values for goods is above a cut-off and the weighted sum of values for the bads is below the cut-off. Since we cannot expect to gain perfect separation we need to add error term variables which are positive or zero to the equation for the goods and subtract them from the equation for the bads. We wish to minimise the sum of the errors over all cases, subject to the above classification rule. Arranging applicants so that the first n_{G} are good and the remaining n_{B} cases are bad, the set up is

$$\begin{aligned}
\text{Min}\quad & a_{i=1} + a_{i=2} + \cdots\cdots\cdots + a_{i=n_{\mathrm{G}}+n_{\mathrm{B}}}\\
\text{s.t.}\quad & w_1x_{1i} + w_2x_{2i} \cdots\cdots w_px_{pi} > c - a_i \quad \text{for } 1 \leqslant i \leqslant n_{\mathrm{G}},\\
& w_1x_{1i} + w_2x_{2i} \cdots\cdots w_px_{pi} \leqslant c + a_i \quad n_{\mathrm{G}} + 1 \leqslant i \leqslant n_{\mathrm{G}} + n_{\mathrm{B}},\\
& a_i \geqslant 0 \quad 1 \leqslant i \leqslant n_{\mathrm{G}} + n_{\mathrm{B}}.
\end{aligned} \tag{6}$$

To avoid the trivial solution of weights and cut-off being set to zero one needs to solve the program once with the cut-off set to be positive and once when negative. But even then the solution may vary under linear transformations of the data. Various modifications have been suggested to overcome this problem. An alternative to the set up of Eq. (6) is to minimise the total cost of misclassifying applicants so the objective function is

$$\text{Min} \quad L\sum_{i=1}^{n_{\mathrm{G}}} m_i + D\sum_{i=n_{\mathrm{G+1}}}^{n_{\mathrm{G+B}}} m_i, \tag{7}$$

where m_i is a dummy taking on values of 1 if a case is misclassified and 0 otherwise and L and D are as above. The constraints are appropriately modified. However this method is feasible only for relatively small samples due to its computational complexity and it is also possible that several different optimal solutions may exist with very different performances on the holdout sample. Work by Glen (1999) has amended the algorithm to avoid some of these problems.

3.3. Neural networks

Neural networks (NN) are a further group of algorithms. Here each characteristic is taken as an ‘input’ and a linear combination of them is taken with arbitrary weights. The structure of a very simple multilayer perceptron is shown in Fig. 4. The central column of circles is a hidden layer; the final circle is the output layer. The characteristics are linearly combined and subject to a non-linear transformation represented by the g and h functions, then fed as inputs into the next layer for similar manipulation. The final function yields values which can be compared with a cut-off for classification (Haykin, 1999). A large choice of activation functions is available (logistic, exponential etc.). Each training case is submitted to the network, the final output compared with the observed value (0 or 1 in this case) and the difference, the error, is propagated back though the network and the weights modified at each layer according to the contribution each weight makes to the error value (the back propagation

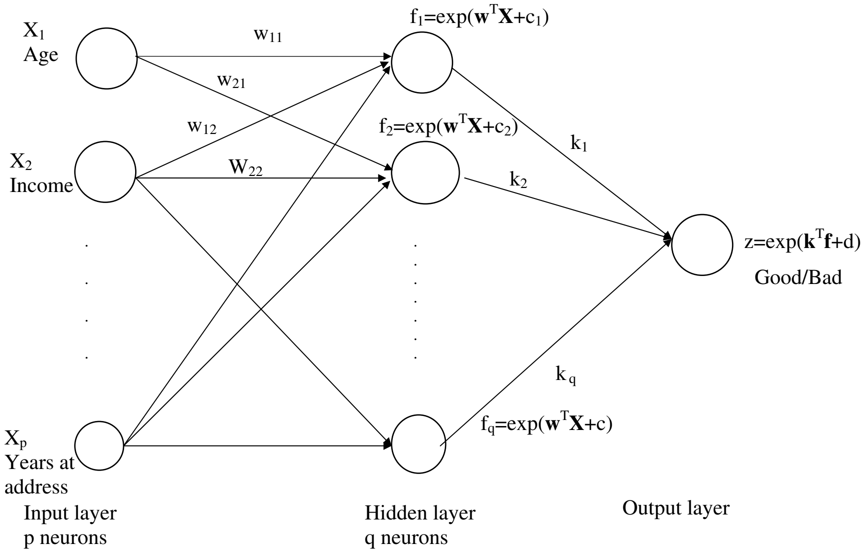

Fig. 4. A two layer neural network.

algorithm). In essence the network takes data in characteristics space, transforms it using the weights and activation functions into hidden value space (the ‘*g*’ values) and then possibly into further hidden value space, if further layers exist, and eventually into output layer space which is linearly separable. In principle only one hidden layer is needed to separate non-linearly separable data and only two hidden layers are needed to separate data where the observations are is in distinct and separate regions. Neural networks have a weakness when used to score credit applicants: the resulting set of classification rules are not easily interpretable in terms of the original input variables. This makes it difficult to use if, as is required in the US and UK, a rejected applicant is given a reason for being rejected. However several algorithms are available that try to derive interpretable rules from a NN. These algorithms are generally of two types: decompositional and pedagogical. The former take an estimated network and prune connections between neurons subject to the difference in predictions between the original network and the new one being within an accepted tolerance. Pedagogical techniques estimate a simplified set of classification rules which is both interpretable and again gives acceptably similar classifications as the original network. Neurolinear is an example of the first type and Trepan an example of the second. Baesens (2003) finds that these approaches can be successful. Of course there is no need to explain the outcomes when the network is being used to detect fraud and whilst networks have been and are used by consultancies for application scoring it is fraud scoring where arguably they have been most commonly used.

3.4. Other classifiers

A similar type of classifier uses fuzzy rather than crisp rules to discriminate between classes. A crisp rule is one which precisely defines a membership criterion (for example “poor” if income is less than $10 k) whereas a fuzzy rule allows membership criteria to be less precisely defined with membership depending possibly on several other less precisely defined criteria. The predictive accuracy of fuzzy rule extraction techniques applied to credit scoring data has been mixed relative to other classifiers. Baesens (2003) finds their performance to be poorer but not significantly so than linear discriminant analysis (LDA), NN and decision trees. But (Piramuthu, 1999) found the performance of fuzzy classifiers to be significantly worse than NN when applied to credit data.

Classification trees are another type of classifier. Here a characteristic is chosen and one splits the set of attributes into two classes that best separate the goods from the bads. Each case is ascribed to one branch of the classification tree. Each subclass is again divided according to the attribute value of a characteristic that again gives the best separation between the goods and bads in that sub-sample or node. The sample splitting continues until few cases remain in each branch or the difference in good rate between the subsets after splitting is small. This process is also called the Recursive Partitioning Algorithm. The tree is then pruned back until the resulting splits are thought to give an acceptable classification performance in the population. Various statistics have been used to evaluate how well the goods and bads are separated at the end of any one branch and so which combination of attribute values to allocate to one branch rather than to the other if the characteristic is a nominal one. One example is the KS statistic described above. Another common statistic advocated by Quinlan (1963) is the entropy index defined as

$$I(\omega) = -P(G|\omega)\ln_2(P(G|\omega)) - P(B|\omega)\ln_2(P(B|\omega)) \tag{8}$$

which relates to the number of different ways one could get the good–bad split in a node and so indicates the information contained in the split. The analyst must decide which splitting rule to use, when to stop pruning, and how to assign terminal nodes to good and bad categories.

Logistic regression is the most commonly used technique in developing scorecards, but linear regression, mathematical programming, classification trees and neural networks have also been used in practice. A number of other classification techniques such as genetic algorithms, support vector machines (SVM) and nearest neighbours have been piloted but have not become established in the industry.

Genetic methods have also been applied to credit data. In a genetic algorithm (GA) a solution consisting of a string (or chromosome) of values (alleles) may represent, for each characteristic, whether a characteristic enters the decision rule and a range of values that characteristic may take. This is represented in binary form. Many candidate solutions are considered and a measure of the predictive accuracy or fitness is computed for each solution. From this sample of solutions a random sample is selected where the probability of selection depends on the fitness. Genetic operators are applied. In crossover the first n digits of one solution are swapped with the last n digits of another. In mutation specific alleles are flipped. The new children are put back into the population. The fitness of each solution is recomputed and the procedures repeated until a termination condition is met. GAs make use of the Schemata theorem (Holland, 1968) that essentially shows that if the average fitness of a schemata is above the average in the population of possible solutions then the expected number of strings in the population with this schemata will increase exponentially. GAs are optimisation algorithms rather than a class of classifier and may be applied as described or to optimise NNs, decision trees and so on. For example when GAs are applied to a decision tree mutation involves randomly replacing the branches from a node in a tree and crossover involves swapping branches from a node in one tree with those in another tree (Ong et al., 2005).

Support vector machines (SVM) can be considered as an extension of the mathematical programming approach but where the scorecard is a weighted linear combination of functions of the variables not just a weighted linear combination of the variables themselves. The objective is for all the goods to have “scores” greater than $c + 1$ and the bads to have “scores” below $c - 1$ so that there is a normalized gap of $2/\|\mathbf{w}\|$ between the two groups. The objective is also extended to minimize a combination of the size of the misclassification errors and the inverse of this gap. In fact, the algorithms work by solving the dual problem when the dual kernel is chosen to be of a certain form. Unfortunately one cannot invert this to obtain the corresponding primal functions and thus an explicit expression for the scorecard.

Many other classifiers may be used to divide applicants into good and bad classes. These include K-nearest neighbour techniques and various Bayesian methods. Some comparisons can be made between the classifiers. LR, LDA, and MP are all linear classifiers: in input space the classifying function is linear. This is not so for quadratic discriminant analysis, where the covariances of the input variables are not the same in the goods and bads, nor in NNs, SVMs or decision trees. Second, all techniques require some parameter selection by the analyst. For example, in LR it is the assumption that a logistic function will best separate the two classes, in NN the number of hidden layers, the types of activation function, the type of error function, the number of epochs and so on; in trees it is the splitting and stopping criteria etc. Third, a common approach to devising a non-linear classifier is to map data from input space into a higher dimensional feature space. This is done with least restrictions in SVMs but is also done in NNs. Furthermore several researchers have experimented with combining classifiers. For example (Schebesch and Stecking, 2005) have used SVMs to select cases (the support vectors) to present to a LDA. Others have used trees to select variables to use in a LR. Yet others have used a stepwise routine in LDA to select the variables for inclusion in a NN (Lee et al., 2002)). Kao and Chiu (2001) have combined decision trees and NNs.

Whilst many comparisons of the predictive performance of these classifiers have been made relatively few have been conducted using consumer credit data and it is possible that the data structures in application data are more amenable to one type of classifier rather than to others. Table 2 gives a comparison of published results, although caution is required since the table shows only the error rate, the papers differ in how the cut-off was set, the errors are not weighted according to the relative costs of misclassifying a bad and a good and few papers use inferential tests to see if the differences are significant. For these reasons the figures can only be compared across a row not between studies. For example Henley's results show a much lower proportion correctly classified because the proportion of defaults in his sample was very different from that in other studies. Nevertheless it would seem from the table that SVMs when included do perform better than other classifiers, a result that is consistent with comparisons using data from other contexts. Neural networks also perform well as we would expect bearing in mind that a NN with a logistic activation function encompasses a LR model. Given that small improvements in classification accuracy results in millions of dollars of additional profits attempts to increase accuracy by devising new algorithms is likely to continue.

However (Hand, 2006) argues that the relative performances illustrated in comparisons like Table 2 overemphasise the likely differences in predictive performance when classifiers are *used in practice*. One reason is associated with sampling issues. Over time "population drift" occurs whereby changes in $p(G|x)$ occur due, in the case of LDA to changes in $p(G)$ or $p(x|G)$, or in the case of LR to changes, not in the parameters of the model, but to changes in the distributions of variables not included in the model but which affect the posterior probability. Another reason is that the training sample may not represent the population of all applicants, as we explain below in Section 4. Further, since the definitions of "good" and "bad" which are used are arbitrary, they may change between the time the scorecard was developed and the time when it is implemented. A model which performs well relative to other classifiers on one definition may perform relatively less well on another. Moreover each classifier typically optimizes a measure of fit (for example maximum likelihood in the case of LR) but its performance in practice may be judged by a very different criterion, for example cost-weighted error rates, for which there are many possibilities. So the relative performances of classifiers may differ according to the fitness measure used in practice.

3.5. Comparisons with corporate risk models

The methods used to predict the probability of default for individuals that we have considered so far may be contrasted with those applied in the corporate sector. One can distinguish between the risk of lending or extending a facility to an individual company and the collective risk of lending to a portfolio of companies. (A corresponding separation applies to loans to individuals). For the former a number of credit rating agencies offer opinions as to the credit worthiness of lending to large quoted companies which are typically indicated by an alphabetic ordinal scale, for example S&P's scale of "*AAA*", "*AA*" … "*CCC*" labels. Different agencies' scales often have different meanings. Moody's scale indicates a view of likely expected loss (including PD and loss given default) of extending a facility whereas S&P's scale reflects an opinion about the PD of an issuer across all of its loans. The details of the methods used to construct a scale are proprietary, however they involve both quantitative analyses of financial ratios and reviews of the competitive environment, including regulations and actions of competitors and also internal factors such as managerial quality and strategies. The senior management of the firm to be rated is consulted and a committee within the agency would discuss its proposed opinion before it is made public.

Although there are monotonic relationships between PDs and ordinal risk categories the risk category is not supposed to indicate a precise PD for a borrower. Some agencies try to produce a classification which is invariant through a business cycle unless a major deterioration in the credit worthiness of the firm is detected, (but in practice some variation through the cycle is apparent (De Servigny and Renault, 2004)), while others give ratings which reflect the point in time credit risk.

An alternative approach, which is typified by KMV's Credit Monitor, is to deduce the probability

Table 2
Relative predictive accuracy of different classifiers using credit application data

Author	Linear regression or LDA	Logistic regression	Decision trees	Math programming	Neural nets	Generic algorithms	Genetic programming	*K*-nearest neighbours	Support vector machines
Percentage correctly classified									
Srinivisan and Kim (1987)	87.5	89.3	**93.2**	86.1					
Boyle et al. (1992)	**77.5**		75.0	74.7					
Henley (1995)[e]	43.4	43.3	**43.8**						
Desai et al. (1997)	66.5	**67.3**			66.4				
Yobas et al. (2000)	**68.4**		62.3		62.0	64.7			
West (2000)[a]	79.3	81.8	77.0		**82.6**			76.7	
Lee et al. (2002)	71.4	73.5			73.7 (**77.0**)[b]				
Malhotra and Malhotra (2003)	69.3				**72.0**				
Baesens, 2003[c]	79.3	79.3	77.0	79.0	**79.4**			78.2	**79.7**
Ong et al., 2005[d]	80.8		78.4		81.7		**82.8**		

[a] Figs are an average across two data sets.

[b] Hybrid LDA and NN.

[c] Figs are averaged across eight data sets. Results are not given for fuzzy extraction rules because benchmark methods had different results compared with those shown in the columns in the table.

[d] Figs are averages over two datasets.

[e] Henley's data had a much higher proportion of defaults than did the other studies.

of default from a variant of Merton's (1974) model of debt pricing. This type of model is called a "structural model". The essential idea is that a firm will default on its debt if the debt outstanding exceeds the value of the firm. When a firm issues debt the equity holders essentially give the firm's assets to the debt holders but have an option to buy the assets back by paying off the debt at a future date. If the value of the assets rises above the value of the debt the equity holders gain all of this excess. So the payoff to the equity holders is like a payoff from a call option on the value of the firm at a strike price of the repayment amount of the debt. Merton makes many assumptions including that the value of the firm follows a Brownian motion process. Using (Black and Scholes, 1973) option pricing model the PD of the bond can be recovered. Whilst in Merton the PDs follow a cumulative normal distribution, $\mathrm{PD} = N^{-1}(\mathrm{DD})$ where DD is distance to default, this is not assumed in KMV Credit Monitor where DD is mapped to a default probability using a proprietary frequency distribution over DD values.

For non-quoted companies the market value of the firm is unavailable and the Merton model cannot be used. Then, corporate risk modelers often resort to methods used to model consumer credit as described in the last few sections.

4. Reject inference

A possible concern for credit risk modelers is that we want a model that will classify new applicants but we can only train on a sample of applicants that were accepted and who took up the offer of credit. The repayment performance of the rejected cases is missing. There are two possible problems here. First, the possibility that the parameters of such a model may be biased estimates of those for the intended population arises. Second, without knowing the distribution of goods and bads in the applicant population one cannot know the predictive performance of the new model on this population.

The first problem is related to Rubin's classification of missing data mechanisms (Little and Rubin, 1987). The two types of missing data mechanisms that are relevant here are missing at random (MAR) and missing not at random (MNAR). In the MAR case the mechanism that causes the data to be missing is not related to the variable of interest. MNAR occurs when the missingness mechanism is closely related to the variable of interest. So if D denotes default, A denotes whether an applicant was accepted, $P(D)$ is to be predicted using predictor variables $\mathbf{X}_1$ and we may try to explain $P(A)$ in terms of $\mathbf{X}_2$, we can write:

$$P(D|\mathbf{X}_1) = P(D|\mathbf{X}_1, A=1)\cdot P(A=1|\mathbf{X}_2) + P(D|\mathbf{X}_1, A=0)\cdot P(A=0|\mathbf{X}_2). \qquad (9)$$

In MAR $P(D|\mathbf{X}_1, A=1) = P(D|\mathbf{X}_1, A=0)$ so $P(D|\mathbf{X}_1) = P(D|\mathbf{X}_1, A=1)$ whereas in MNAR $P(D|\mathbf{X}_1, A=1) \neq (D|\mathbf{X}_1, A=0)$ and the sample of accepted applicants will give biased estimates of the population parameters. Hand and Henley (1994) have shown that if the $\mathbf{X}_1$ set encompasses the original variables used to accept or reject applicants we have MAR, otherwise we have MNAR. If we have MAR then a classifier like LR, NNets, and SVMs which do not make distributional assumptions will yield unbiased estimates.

Researchers have experimented with several possible ways of incorporating rejected cases in the estimation of a model applicable to all applicants. These have included the EM algorithm (Feelders, 2000), and multiple imputation, but these together with imputation by Markov Chain Monte Carlo simulation have usually assumed the missingness mechanism is MAR when it is not known if this is the case. If it is the case then such procedures will not correct for bias since none would be expected. However if we face MNAR then we need to model the missingness mechanism. An example of this procedure is Heckman's two stage sample selection model (Heckman, 1979).

In practice a weighting procedure known as augmentation (or variants of it) is used. Augmentation involves two stages. In the first stage one estimates $P(A|\mathbf{X}_2)$ using an ex post classifier and the sample of all applicants including those rejected. $P(A|\mathbf{X}_2)$ is then predicted for those previously accepted cases and each is ascribed a "sampling" weight of $1/P(A|\mathbf{X}_2)$ in the final LR. Thus each accepted case represents $1/P(A|\mathbf{X}_2)$ cases and it is assumed that at each $P(A|\mathbf{X}_2)$ value $P(D|\mathbf{X}_1, A=1) = P(D|\mathbf{X}_1, A=0)$. That is, the proportion of goods at that value of $P(A|\mathbf{X}_2)$ amongst the accepts equals that amongst the rejects, and so the accepts at this level of $P(A|\mathbf{X}_2)$ can be weighted up to represent the distribution of goods and bads at that level of $P(A|\mathbf{X}_2)$ in the applicant population as a whole.

An alternative possible solution is to use a variant of Heckman's sample selection method whereby one estimates a bivariate probit model with selection (BVP) which models the probability of default and the probability of acceptance allowing for the fact that observed default is missing for those not accepted (Crook and Banasik, 2003). However this method has several weaknesses including the assumption that the residuals for the two equations are bivariate normal and that they are correlated. The last requirement prevents its use if the original accept model can be perfectly replicated.

There is little empirical evidence on the size of this possible bias because its evaluation depends on access to a sample where all those who applied were accepted and this is very rarely done by lenders. However a few studies have access to a sample which almost fits this requirement and they find that the magnitude of the bias and so scope for correction is very small. Experimentation has shown that the size of the bias increases monotonically if the cut-off used on the legacy scoring model increases and the accept rate decreases (Crook and Banasik, 2004). Using the training sample rather than an all applicant sample as holdout leads to a considerable overestimate of a model's performance, particularly at high original cut-offs. Augmentation has been found to actually reduce performance compared with using LR for accepted cases only. Recent work, some of which is reported on in this issue, investigates combining both the BVP and augmentation approaches, altering the BVP approach so that the restrictive distributional assumptions about the error terms are relaxed and uses branch and bound algorithms.

5. Profit scoring

Since the mid 1990s a new strand of research has explored various ways in which scorecards may be used to maximize profit. Much of this work is proprietary, but many significant contributions to the literature have also been made. One active area concerns cut-off strategy. In a series of path breaking papers Oliver and colleagues (Marshall and Oliver, 1993; Hoadley and Oliver, 1998; Oliver and Wells, 2001) have formulated the problem as follows. For a given scoring model one can plot the expected bad rate (i.e. the expected proportion of all accepts that turn out to be bad or $(1 - F(s))p_{\mathrm{B}}$ where $F(s)$ is the proportion of applicants with a score less than s) against the acceptance rate (i.e. the proportion of applicants that are accepted or $(1 - F(s))$). This is shown in Fig. 5 and is called a strategy curve. A perfectly discriminating model will, as the cut-off score increases, accept all goods amongst those that apply and then accept only bads. In Fig. 5 if it is assumed that 95% of applicants are good, the perfect scorecard is represented by the straight line along the horizontal axis until the accept rate is 95%, thereafter the curve slopes up at 45°. A replacement scoring model may be represented by the dotted line. The bank may currently be at point P and the bank has a choice of increasing the acceptance rate whilst maintaining the current expected bad rate (A) or keeping the current acceptance rate and decreasing the expected bad rate (B), or somewhere in between. The coordinates of all of these points can be calculated from parameters yielded by the estimation of the scorecard.

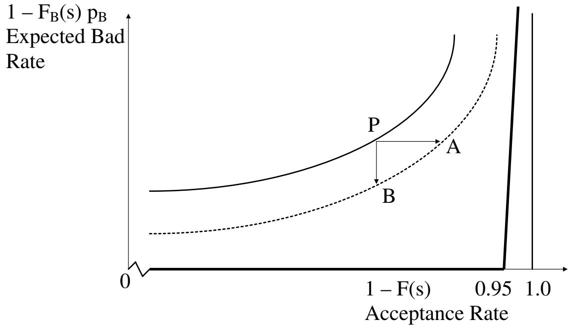

Fig. 5. The strategy curve.

One can go further and plot expected losses against expected profits (Fig. 6). For a given scoring model as one lowers the cut-off an increasing proportion of accepts are bad. Given that the loss from an accepted bad is many times greater than the profit to be made from a good, this reduction in cut-off at first increases profits and the expected loss, but eventually the increase in loss from the ever increasing proportion of bads outweighs the increase in profits and profits fall. Now, the bank will wish to be on the lower portion of the curve

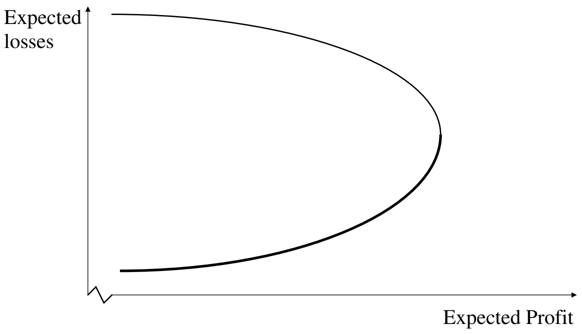

Fig. 6. The efficiency curve.

which is called the efficiency frontier. The bank may now find its current position and, by choice of cut-off and adoption of the new model, decide where on the efficiency frontier of a new model it wishes to locate: whether to maintain expected losses and increase expected profits or maintain expected profit and reduce expected losses. Oliver and Wells (2001) derive a condition for maximum profits which is, unsurprisingly that the odds of being good equal the ratio of the loss from a bad divided by the gain from a good (D/L). Interestingly, an expression for the efficiency frontier can be gained by solving the non-linear problem of minimizing expected loss subject to expected profits being at least a certain value P_0. The implied cutoff is then expressed as $H^{-1}\frac{P_0}{L\cdot p_g}$ where H is an expression for expected profits and all values can be derived from a parameterized scoring model.

So far the efficiency frontier model has omitted interest expense on funds borrowed, operating costs and the costs of holding regulatory capital as required under Basel 2. In ongoing work the model has been amended by Oliver and Thomas (2005) by incorporating these elements into the profit function. They show that for a risk neutral bank the requirements of Basel 2 unambiguously increase the cut-off score for any scorecard and so reduce the number of loans made compared with no minimum capital requirement.

Some recent theoretical models of bank lending have turned their attention to more general objectives of a bank than merely expected profit. For example Keeney and Oliver (2005) have considered the optimal characteristics of a loan offer from the point of view of a lender, given that the utility of the lender depends positively on expected profit and on market share. The lender's expected profit depends on the interest rate and on the amount of the loan made to a potential borrower. To maximize this utility a lender has to consider the effects of both arguments on the probability that a potential borrower will take a loan offered and so on market share. The utility of a potential borrower is assumed to depend positively on the loan amount and negatively on the interest rate. It is also crucially assumed that as the interest rate and amount borrowed increase, whilst keeping the borrower at the same level of utility and so market share, the expected profits of the lender increase, reach a maximum and then decrease. By comparing the maximum expected profit at each utility level the lender can work out the combination of interest and loan amount to offer to minimize its expected profit. Because levels of the interest rate have the opposite effects for a borrower and a lender, as one increases this rate and reduces the loan amount the probability of acceptance and so market share goes down whilst at first expected profits rise. The lender has to choose the combination of market share and expected profit it desires and then compute the interest rate and loan amount to offer to the applicant to achieve this. Notice however some limitations of this model. First it assumes that there is only one lender, since the probability of acceptance is not made dependent on the offers made by rival lenders. Second, the model omits learning by both lender and potential borrower. Third, this is very difficult to implement in practice since in principle one would need the utility function for every individual potential borrower.

This and associated models has spawned much new research. For example some progress has been made on finding the factors that affect the probability of acceptance. Seow and Thomas (2007) have presented students with hypothetical offers of a bank account and each potential borrower had a different set of (hypothetical) account characteristics offered to them. The aim was to find the applicant and account characteristics that maximized acceptance probability. Unlike earlier work, by using a classification tree in which offer characteristics varied according to the branch the applicant was on, this paper does not assume applicants have the same utility function.

6. The Basel 2 accord

A major research agenda has developed from the Basel 2 Accord. The Basel Committee on Banking and Supervision was set up in 1975 by the Governors of the G10 most industrialised countries. It has no formal supervisory authority but it states supervisory guidelines and expects legal regulators in major economies to implement them. The aims of the guidelines were to stabilize the international banking system by setting standardised criteria for capital levels and a common framework to facilitate fair banking competition. The need for such standards was perceived after major reorganisations in the banking industry, certain financial crises and bank failures, the very rapid development of new financial products and increased banking competition.

The first Basel accord (Basel, 1988), stated that a bank's capital had to exceed 8% of its risk weighted assets. But the accord had many weaknesses (see De Servigny and Renault, 2004) and over the last seven years the BIS has issued several updated standards which have come to be known as the Basel 2 accord. It is due to be adopted from 2007-08. (Pillar 3 does not take effect until later.)

The Basel 2 Accord (Basel, 2005) states that sound regulation consists of three 'pillars': minimum capital requirements, a Supervisory Review Process and market discipline. We are concerned with the first. Each bank has a choice of method it may use to estimate values for the capital it needs to set aside. It may use a Standardised Approach in which it has to set aside 2.4% of the loan capital for mortgage loans and 6% for regulatory retail loans (to individuals or SMEs for revolving credit, lines of credit, personal loans and leases). Alternatively the bank may use Internal Ratings Based approaches, in which case parameters are estimated by the bank. Notice that if a bank adopts an IRB for, say, its corporate exposures it must do so for all types of exposures. The attraction of the IRB approach over the Standardised Approach is that the former may enable a bank to have less capital and so earn higher returns on equity than the latter. Therefore it is likely that most large banks will adopt the IRB approach.

Under the IRB approach a bank must retain capital to cover both expected and unexpected losses. Expected losses are to be built into the calculation of profitability that underlie the pricing of the product, and so covered by these profits. The unexpected losses are what the regulatory capital seeks to cover. For each type of retail exposure that is not in default the regulatory capital requirement, k, is calculated as

$$k = \mathrm{LGD} \times N\left[\frac{1}{\sqrt{1-R}} \cdot N^{-1}(\mathrm{PD}) + \sqrt{\frac{R}{1-R}} \cdot N^{-1}(0.999)\right] - \mathrm{PD} \times \mathrm{LGD}, \qquad (10)$$

where LGD is loss given default (the proportion of an exposure which, if in default, would not be recovered); N is the cumulative distribution for a standard normal random variable; N^{-1} is inverse of N; R is the correlation between the performance of different exposures; and PD is the probability of default. R is set at 0.15 for mortgage loans, at 0.04 for revolving, unsecured exposures to individuals up to €100 k, and

$$R = 0.03 \times \left[\frac{(1-\exp(-35 \times \mathrm{PD})}{(1-\exp(-35))}\right] + 0.16 \times \left[1 - \left(\frac{1-\exp(-35 \times \mathrm{PD})}{1-\exp(-35)}\right)\right] \qquad (11)$$

for other retail exposures.

Intuitively the second term in Eq. (10) (involving the square brackets) represents the probability that the PD lies between its predicted value and its 99.9th percentile value. Its formal derivation is given in Schonbucher (2000) and Perli and Nayda (2004). The formula is derived by using the assumption that a firm will default when the value of its assets falls below a threshold which is common for all firms. The value of the firm is standard normally distributed and, crucially, the assets of all obligors depend on one common factor, and a random element which is particular to each firm. The third term represents future margin income which could be used to reduce the capital required. This is included because the interest on each loan includes an element to cover for losses on defaulted loans and this can be used to cover losses before the bank would need to use its capital. The size of this income is estimated as the expected loss.

For each type of exposure the bank has to divide its loans into 'risk buckets' i.e. groups which differ in terms of the group's PD but within which PDs are similar. When allocating a loan to a group the bank must take into account borrower risk variables (sociodemographics etc.), transaction risk variables (aspects of collateral) and whether the loan is currently delinquent. To allocate loans into groups behavioural scoring models may be used with possible overrides. The PD value for each group has to be valid over a five year period to cover a business cycle. For retail exposures several methods to estimate the PD of each pool are available but most banks will use a behavioural scoring model.

The Basel 2 Accord has led to many new lines of research. The first issue is the appropriateness of the one factor model to estimate unexpected loss of a risk group for consumer debt since this model was originally devised for corporate default. Although Perli and Nayda (2004) build such a consumer model several authors have questioned the plausibility of a household defaulting on unsecured debt if the value of the household's assets dip below a threshold. Instead new models involving reputational assets and jump process models have been tried. De Andrade and Thomas (2007) assume that a borrower will default when his reputation dips below a threshold. His reputation is a function of a latent unobserved variable representing his credit worthiness.

A second topic of research is the assumed parameter values in the Basel formula, specifically the assumed correlation values. De Andrade and Thomas found that the calculated correlation coefficient for 'other retail exposures' for a sample of Brazilian data was 2.28% which is lower than the correlation range of 3% in Basel 2. Lopez (2002) found the correlation between asset sizes to be negatively related to default probability for firms, which is the opposite of the relationship implied by Eq. (11), and to be positively related to firm size which is not included in Eq. (11).

A third source of research is how to estimate PDs from portfolios with very low default proportions. In very low risk groups, especially in mortgage portfolios, very few, sometimes no defaults, are observed. Estimating the average PD in such a group as close to zero would not reflect the conservatism implied by Basel 2. Pluto and Tasche (2005) assume the analyst knows at least the rating grade into which a borrower fits and calculates confidence intervals so that by adopting a 'most prudent estimation principle' PD estimates can be made which are monotone in the risk groups. For example consider the case of no defaults. Three groups, $i = 1,2,3$ in decreasing order of risk each have n_1, n_2 and n_3 cases respectively. The group risk ordering implies $P_1 \leqslant P_2 \leqslant P_3$. The most prudent estimate of P_1 implies $P_1 = P_2 = P_3$. Assuming independence of defaults the probability of observing zero defaults is $(1 - P_1)^{n1+n2+n3}$ and the confidence region at level γ for P_1 is such that $P_1 \leqslant 1 - (1-\gamma)^{n1+n2+n3}$. At a 99.9% confidence level, $P_1 = 0.86\%$. By similar reasoning the most prudent value of P_2 is $P_2 = P_3$ and the upper boundary may be calculated (at 99.9% it is 0.98%). The difference in values is a consequence of the number of cases. In the case of a few defaults and assuming the number of defaults follows a binomial distribution, the upper confidence interval can also be calculated. Pluto and Tasche extend the calculations using a one factor probit model to find the confidence interval for the PD when asset returns are correlated. This increases the PDs considerably and the new probabilities under the same assumptions as above were $P_1 = 5.3\%$, $P_2 = 5.8\%$ and $P_3 = 9.8\%$.

Forrest (2005) adopted a similar approach of calculating the 95%ile of the default distribution. But unlike Pluto and Tasche he calculated the PD as the value that maximises the likelihood of observing the actual number of defaults. So for a portfolio consisting of two grades, A with a higher PD than B and d_a defaults and $(1 - d_a)$ exposures, and B with d_b defaults and $(1 - d_b)$ exposures, and letting P_{A} and P_{B} indicate their respective PDs, P_{A} and P_{B} can be calculated by choosing their values to maximise

$$\text{Likelihood}\; L(P_{\mathrm{A}}, P_{\mathrm{B}}) = P_{\mathrm{A}}^{d_a}(1 - P_{\mathrm{A}})^{1-d_a} P_{\mathrm{B}}^{d_b}(1 - P_{\mathrm{B}})^{1-d_b} \tag{12}$$

constrained so that $P_{\mathrm{A}} > P_{\mathrm{B}}$.

A further issue is how to validate estimates of PDs, LGDs and EADs. The application and behavioural scoring models outlined in Section 2 are required to provide only an ordinal ranking of cases whereas ratio level values are required for Basel 2 compliance. Validation of PDs has two aspects; discriminatory power, which refers to the degree of separation of the distributions of scores between groups, and calibration which refers to the accuracy of the PDs. We considered the former in an earlier section; here we consider the latter. The crucial issue is how to test whether the size of the difference between estimated *ex ante* PDs and *ex post* observed default rates is significant. Several tests have been considered by the BIS Research Task Force (2005) (BISTF) as follows. The test is required for each risk pool. The binomial test, which assumes the default events within the risk group are independent, tests the null hypothesis the PD is underestimated. Although the BISTF argue that empirically estimated correlation coefficients between defaults within a risk group are around 0.5–3% they assess the implications of different asset correlations by reworking the one factor model with this assumption. They find that when the returns are correlated, an unadjusted binomial test will be too conservative: it will reject the null when the correct probability of doing so is in fact much lower than intended. The BISTF reports that a chi-square test could be used to test the same hypotheses as the binomial test but is applied to the portfolio as a whole (i.e. to all risk groups) to circumvent the well known problem that multiple tests would be expected to indicate a rejection of the null in 5% of cases. But again it has the same problems as the binomial test when defaults are correlated. The normal test allows for correlated defaults and uses the property that the distribution of the standardised sum of differences between the observed default rate in each year and the predicted probability is normally distributed. The null is that 'none of the true probabilities of default in the years $t = 1, 2, \ldots, T$ is greater than the forecast P_t'. However the BISTF find the test has conservative bias: in too many cases the null is rejected. A further test is proposed by Blochwitz et al. (2004) and called the 'traffic lights test'. Ignoring dependency in credit scores over t, the distribution of the number of defaults is assumed to follow a multinomial distribution over four categories labelled 'green', 'yellow', 'orange', and 'red'. The probability that a linear combination of the number of defaults in each category exceeds a critical value is used to test the null that 'none of the true probabilities of default in years $t = 1, 2, \ldots, T$ is greater than the corresponding forecast PD_t'. Running simulation tests the BISTF find the test is rather conservative.

After comparisons of these four inferential test statistics the BISTF concluded that currently 'no really powerful tests of adequate calibre are currently available' and call for more research.

7. Conclusion

In this paper we have reviewed a selection of current research topics in consumer credit risk assessment. Credit scoring is central to this process and involves deriving a prediction of the risk associated with making a loan. The commonest method of estimating a classifier of applicants into those likely to repay and those unlikely to repay is logistic regression with the logit value compared with a cut off. Many alternative classifiers have been developed in the last thirty years and the evidence suggests that the most accurate is likely to be support vector machines although much more evidence is needed to confirm this tentative view. Behavioural scoring involves the same principles as application scoring but now the analyst has evidence on the behaviour of the borrower with the lender and the scores are estimated after the loan product has been taken and so updated risk estimates can be made as the borrower services his account. Initial concerns that the estimation of classifiers based on only accepted applicants and not on a random sample of the population of applicants seem to be rather over placed, but using applicants only to assess the predictive performance of a classifier does appear likely to lead to significant over optimism as to a predictor's performance. Scoring for profit is a major research initiative and has led to some important tools which have been employed by consultancies over the last few years. Arguably the most significant factor impacting on consumer credit scoring processes is the passing of the Basel 2 accord. This has determined the way in which banks in major countries calculate their reserve capital. The amount depends on the use of updated scoring models to estimate the probability of default and other methods to estimate the loss given default and amount owing at default. The method is based on a one factor Merton model but has been subject to major criticisms. The effect of Basel 2 on the competitive advantage of an institution is so large that research into the adequacy, applicability and validity of the adopted models is set to be a major research initiative in the foreseeable future.

There are many topics that we do not have space in this paper to consider (see Thomas et al. (2002) for further details). Consumer credit scoring is also used to decide to whom to make a direct mailshot, it is used by government agencies (for example tax authorities) when deciding who is likely to pay debts, it is used by utilities when deciding whether to allow consumers to use electricity/gas etc. in advance of payment. A major use of scoring techniques is in assessing small business loan applications and their monitoring for credit limit adjustment, but this topic alone would fill a complete review paper.

Credit scoring and risk assessment has been one of the most successful applications of statistical and operational research concepts ever with considerable social implications. It has made practical the assessment and granting of hundreds of millions of credit card applications and applications for other loan products and so it has considerably improved the lifestyles of millions of people throughout the world. It has considerably increased competition in credit markets and arguably reduced the cost of borrowing to many. The techniques developed have been applied in a wide variety of decision making contexts so reducing costs to those who present minimal risk to lenders.

References

Anderson, T.W., Goodman, L.A., 1957. Statistical inference about Markov chains. Annals of Mathematical Statistics 28, 89–110.

Baesens, B., 2003. Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques, Doctoral Thesis no 180 Faculteit Economische en Toegepaste Economische Wetebnschappen, Katholieke Universiteit, Leuven.

Basel, 1988. Basel Committee on Banking Supervision: International Convergence of Capital Measurement and Capital Standards. Bank for International Settlements, July 1988.

Basel, 2005. Basel Committee on Banking Supervision: International Convergence of Capital Measurement and Capital Standards: A Revised Framework. Bank for International Settlements, November 2005.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. Journal of Political Economy 81, 637–659.

Blochwitz, S., Hohl, S.D., Tasche, D., When, C.S., 2004. Validating Default Probabilities in Short Times Series. Mimeo, Deutsche Bundesbank.

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.E. (Eds.), Credit Scoring and Credit Control. Oxford University Press, Oxford.

Crook, J.N., Banasik, J., 2003. Sample selection bias in credit scoring models. Journal of the Operational Research Society 54, 822–832.

Crook, J., Banasik, J., 2004. Does reject inference really improve the performance of application scoring models? Journal of Banking and Finance 28 857–874.

Cyert, R.M., Thompson, G.L., 1968. Selecting a portfolio of credit risks by Markov chains. The Journal of Business 1, 39–46.

De Andrade, F.W.M., Thomas, L.C., 2007. Structural models in consumer credit. European Journal of Operational Research, this issue, doi:10.1016/j.ejor.2006.07.049.

Desai, V.S., Conway, 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.

De Servigny, A., Renault, O., 2004. Measuring and Managing Credit Risk. McGraw Hill, New York.

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

Feelders, A.J., 2000. Credit scoring and reject inference with mixture models. International Journal of Intelligent Systems in Accounting, Finance and Management 9, 1–8.

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

Forrest, A., 2005. Low Default Portfolios – Theory and Practice. Presentation to Credit Scoring and Credit Control 9 conference, Edinburgh, 7–9 September 2005.

Frydman, H., 1984. Maximum likelihood estimation in the Mover–Stayer model. Journal of the American Statistical Association 79, 632–638.

Glen, J., 1999. Integer programming models for normalisation and variable selection in mathematical programming models for discriminant analysis. Journal of the Operational Research Society 50, 1043–1053.

Hand, D.J., 2005. Good practice in retail credit scorecard assessment. Journal of the Operational Research Society 56 (9), September.

Hand, D.J., 2006. Classifier technology and the illusion of progress. Statistical Science 21, 1–14.

Hand, D.J., Henley, W.E., 1994. Inference about rejected cases in discriminant analysis. In: Diday, E., Lechvallier, Y., Schader, P., Bertrand, P., Buntschy, B. (Eds.), New Approaches in Classification and Data Analysis. Springer-Verlag, Berlin, pp. 292–299.

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

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Prentice Hall, New-Jersey.

Heckman, J.J., 1979. Sample selection bias as a specification problem. Econometrica 47, 153–161.

Henley, W.E., 1995. Statistical Aspects of Credit Scoring. Ph.D. thesis, Open University.

Hoadley, B., Oliver, R.M., 1998. Business measures of scorecard benefit. IMA Journal of Mathematics Applied in Business and Industry 9, 55–64.

Ho, J., Thomas, L.C., Pomroy, T.A., Scherer, W.T., 2004. Segmentation in Markov Chain consumer credit behavioural models. In: Thomas, L.C., Edelman, D., Crook, J.N. (Eds.), Readings in Credit Scoring. Oxford University Press, Oxford.

Holland, J.H., 1968. Hierarchical Description of Universal and Adaptive Systems, Department of Computer and Communications Sciences. University of Michigan, Ann Arbor.

Kao, L.J., Chiu, C.C., 2001. Mining the customer credit by using the neural network model with classification and regression. In: Joint 9th IFSA World Congress and 20th NAFIPS International Conference, Proceedings, vols. 1–5, pp. 923–928.

Keeney, R.L., Oliver, R.M., 2005. Designing win–win financial loan products for consumers and businesses. Journal of the Operational Research Society 56 (9), September.

Lee, T.S., Chiu, C.C., Lu, C.J., Chen, I.F., 2002. Credit scoring using the hybrid neural discriminant technique. Expert Systems with Applications 23, 245–254.

Lewis, E., 1992. Introduction to Credit Scoring. The Athena Press, San Rafael.

Little, R.J., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.

Lopez, J., 2002. The Empirical Relationship Between Average Asset Correlation, Firm Probability Of Default And Asset Size. Working Paper, Federal Reserve Bank of San Francisco.

Malhotra, R., Malhotra, D.K., 2003. Evaluating consumer loans using neural networks. Omega 31, 83–96.

Marshall, K.T., Oliver, R.M., 1993. Decision Making and Forecasting. McGraw-Hill, New York, pp. 121–128.

Oliver, R.M., Thomas, L.C., 2005. How Basel Will Affect Optimal Cut-offs Basel Capital Requirements. Presented at Credit Scoring and Credit Control 9 Conference, Edinburgh, September.

Oliver, R.M., Wells, E., 2001. Efficient frontier cut-off policies in credit portfolios. Journal of Operational Research Society 52, 1025–1033.

Ong, C.S., Huang, J.J., Tzeng, G.H., 2005. Building credit scoring systems using genetic programming. Expert Systems with Applications 29, 41–47.

Perli, R., Nayda, W.I., 2004. Economic and regulatory capital allocation for revolving retail exposures. Journal of Banking and Finance 28, 789–809.

Piramuthu, S., 1999. Financial credit-risk evaluation with neural and neurofuzzy systems. European Journal of Operational Research 112, 310–321.

Pluto, K., Tasche, D., 2005. Estimating Probabilities for Low Default Portfolios. Presented at Credit Scoring and Credit Control 9 Conference, Edinburgh, 2005.

Quinlan, J., 1963. C4. 5 Programs for Machine Learning. Morgan-Kaufman, San Mateo, CA.

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

Schebesch, K.B., Stecking, R., 2005. Support vector machines for classifying and describing credit applicants: Detecting typical and critical regions. Journal of the Operational Research Society 56, 1082–1088.

Schonbucher, P.J., 2000. Factor Models For Portfolio Credit Risk. Mimeo, Department of Statistics, Bonn University, December 2000.

Seow, H.-V., Thomas, L.C., 2007. To ask or not to ask, that is the question. European Journal of Operational Research, this issue, doi:10.1016/j.ejor.2006.08.061.

Srinivisan, V., Kim, Y.H., 1987. Credit granting a comparative analysis of classificatory procedures. Journal of Finance 42, 655–683.

Tasche, D., 2005. Rating and probability of default validation in BIS Research Task Force (2005) Basel Committee on Banking Supervision: Studies on the Validation of Internal Rating Systems. Bank for International Settlements, Working Paper No. 14, May 2005.

Thomas, L.C., 2000. A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. International Journal of Forecasting 16, 149–172.

Thomas, L.C., Edelman, D.B., Crook, J.N., 2002. Credit Scoring and its Applications. SIAM, Philadelphia.

Till, R., Hand, D., 2001. Markov Models in Consumer Credit. Presented at Credit Scoring and Credit Control 7 Conference, Edinburgh, September.

Trench, M.S., Pederson, S.P., Lau, E.T., Ma, L., Wang, H., Nair, S.K., 2003. Managing credit lines and prices for bank one credit cards. Interfaces 33 (5), 4–21.

West, D., 2000. Neural network credit scoring models. Computers and Operational Research 27, 1131–1152.

Yobas, M.B., Crook, J.N., Ross, P., 2000. Credit scoring using neural and evolutionary techniques. IMA Journal of Mathematics Applied in Business and Industry 11, 111–125.