Journal of the Operational Research Society (2005) 56, 1109–1117

www.palgrave-journals.com/jors

Good practice in retail credit scorecard assessment

DJ Hand*

Imperial College, London, UK

In retail banking, predictive statistical models called 'scorecards' are used to assign customers to classes, and hence to appropriate actions or interventions. Such assignments are made on the basis of whether a customer's predicted score is above or below a given threshold. The predictive power of such scorecards gradually deteriorates over time, so that performance needs to be monitored. Common performance measures used in the retail banking sector include the Gini coefficient, the Kolmogorov–Smirnov statistic, the mean difference, and the information value. However, all of these measures use irrelevant information about the magnitude of scores, and fail to use crucial information relating to numbers misclassified. The result is that such measures can sometimes be seriously misleading, resulting in poor quality decisions being made, and mistaken actions being taken. The weaknesses of these measures are illustrated. Performance measures not subject to these risks are defined, and simple numerical illustrations are given.
Journal of the Operational Research Society (2005) **56,** 1109–1117. doi:10.1057/palgrave.jors.2601932
Published online 2 February 2005

Keywords: credit scoring; scorecard; bad rate; banking; management; methodology; decision making

Introduction

This paper is concerned with decision making in retail banking. In particular, it is concerned with choosing actions, which are appropriate for individual customers. We assume we have a retrospective database which contains descriptive characteristics of previous customers and also includes aspects of their subsequent behaviour. This information is used to construct a model which will permit us to predict the probable behaviour of new customers on the basis of their descriptive characteristics, so that, if necessary, an appropriate intervention can be made. Many such models have been developed for this problem in the retail credit industry.[1–6] We shall refer to these tools as *scorecards*, in accordance with the terminology used in the consumer credit industry.

The model will predict into which behaviour class a customer is likely to fall, so that we can take an appropriate action. Different models will be built for each aspect of behaviour we wish to predict, of course. For simplicity, in this paper we will assume that, whatever the aspect of behaviour in which we are interested, there are just two possible behaviour classes, each with a unique outcome, and just two corresponding actions which may be taken. The two class case is overwhelmingly the most important, and has become something of a paradigm for the industry. For example, customers are assumed to be either 'good' or 'bad' and logistic prediction models are commonly used. Familiar examples of behaviour class pairs in consumer credit are (default, non-default), (churn, not churn), and (take up offer, do not take up offer). The corresponding actions might be (do not offer loan, offer loan), (attempt to induce the customer to stay, do not incur this expense), and (offer a certain product, do not make this offer), respectively. This restriction to just two classes of customer, with two possible actions, means that we will not be discussing developments such as profitability scorecards. Using this framework, which is very widely used in retail banking, this paper is concerned with criteria for assessing the performance of such scorecards. These performance assessments will be used to choose between alternative possible scorecards, and to monitor scorecard performance over time to decide when the predictive power has deteriorated to the extent that the scorecard needs replacing by a new one.

Formally, we can describe the scorecard as providing a mapping, from the data describing the customers, to the binary space of the two classes. Unfortunately, this mapping is not infallible—we cannot always correctly predict the class into which the customer will eventually fall. This is precisely why we need scorecard assessment criteria.

The mapping from the information describing the customer to the binary class space is typically conducted in two stages. Firstly, the available information is combined to yield a numeric score such that low scores indicate that a customer is more likely to belong to one class and high scores that they are more likely to belong to the other. Secondly, the score is compared to a threshold, t, with customers scoring below the threshold being assigned to one class and those scoring above the threshold being assigned to the other class. For convenience, we will take the high scoring class as the 'good' customers and the low scoring class as the 'bad' customers, but this is easily adjusted for scorecards which score in the opposite direction. The term

*Correspondence: DJ Hand, Department of Mathematics, Imperial College, London, SW7 2AZ, UK.
E-mail: d.j.hand@imperial.ac.uk

'scorecard' arises because of the numerical score produced in the first stage, even though, in situations when an action has to be taken, this is merely a step on the way towards the final assignment to one of the two classes. Perhaps we should remark here that there are one or two exceptions to this two stage strategy. For example, some implementations of neural networks and support vector machines simply yield a final binary classification, without explicitly reporting the intermediate score stage. We should also make explicit the fact that the score is monotonically increasingly related to an estimate of the probability that a customer belongs to the 'good' class. Sometimes (as we shall see below) it is more convenient to work with scores which are actual probability estimates (and, in particular, lie between 0 and 1). Any numerical score can be converted to such a scale (can be calibrated) by a straightforward transformation derived from the data.

Fundamental to the development which follows is the assumption that each behaviour class has a single outcome, and that each of these is associated with an appropriate action which may be taken. This assumption has critical implications for the nature of the assessment criterion used. In particular, it means that only the *sign* of the difference between the score and the threshold matters: if a score exceeds the threshold, then it does not matter by how much, since the decision and, more importantly, the *action*, will be the same whether the score exceeds the threshold by a large amount or a small amount. However, as we shall see, some criteria in widespread use in the industry are not based solely on whether or not the score exceeds the threshold, but also on the extent to which the score differs from the threshold. The use of this extra, irrelevant, information can sometimes lead to seriously misleading conclusions about how well the scorecard is performing, with the result that poor decisions could be taken.

We need to distinguish, at the start, between two different kinds of situation which may occur. The first situation arises with application scorecards. These are used to make an accept/reject decision, in which 'accept' means that the customer is granted a product (a loan, for example) and 'reject' means that they are not. In such situations, the true class—whether they turn out to be good (ie, do not default on the loan) or bad (ie, do default) is eventually observed only for the accepts, and is not observed for those not given a loan. This, of course, is the familiar problem considered in reject inference.[7–9] The consequence is that the performance criterion cannot make use of the (unknown) true classes of the rejected applicants.

The second situation arises when a scorecard is being used to monitor a portfolio of customers; for example, to predict which of them may go bad. In this case, there is no notion of rejection, but a customer identified as having a high probability of going bad will be allocated to a particular procedure (eg, perhaps restrictions might be imposed on their borrowing, or some sort of encouragement to repay might be communicated). Now customers from classes on both sides of the threshold can be used in the performance measure. This second situation can, of course, be generalized to more than two classes. Thus, we could split the range of estimated probability of going bad into groups, with different groups being assigned to different operational procedures. In principle, such extensions are straightforward generalizations of the case of a single split, though some ingenuity may be required to develop criteria for multigroup cases.

In practical implementations, there are issues which we have not mentioned above. Firstly, since the true classes of customers are not discovered until some time in the future, all scorecards are out of date as soon as they are implemented, in terms of how well they match the current population of customers or applicants. Secondly, illustrating with the example of a loan, one cannot be certain that a customer is good until the end of the loan term—they could go bad at any time right up until the end. Of course, one can be sure that they are *bad* before the term if they do actually default. Thirdly, having identified that an existing customer has a high probability of going bad, one will generally adopt some kind of remedial action, and this may (indeed, one hopes it will) influence the eventual outcome. This means that the final proportion going bad will be different from the predicted proportion going bad, so that one cannot use the difference between these proportions to assess the accuracy of the scorecard predictions. This is similar to the problem of reject inference. We do not dwell on such problems here simply because they are universal problems, and are not specific to the tools and methods described here.[10–12]

The next section describes measures for use when one of the actions is 'reject', so that outcomes cannot be observed for this action, and the two subsequent sections describe measures for use when the outcome can be observed under both actions. The penultimate section briefly discusses some other important issues, and the final section draws some conclusions.

When one action is rejection

This case arises in application scoring, where people will be granted or not granted a product on the basis of their score on an application scorecard; for example, a loan. As we noted in the previous section, one fundamental difficulty is that the true, good or bad, class will (later) be observed for those offered the product (those whose score is above the threshold t) but not for those not offered the product (those whose score is below t).

When only the sign of the difference between the score and the threshold, but not the magnitude of this difference, matters, all of the relevant information is contained in the counts of applicants falling above and below the threshold. That is, the number of applicants who score above t and who do not default (ie, who turn out to be good), the number of

applicants who score above t and who do default (ie, who turn out to be bad), and the number of applicants who score below t. The third of these, the number who score below t provides no information about the ability of the scorecard to separate goods from bads though it may play a role in choosing t. Similarly, the absolute number who score above t has no bearing on the measure of separation between the two groups: only the relative numbers, or proportions, of goods and bads above t are relevant. (Of course, if significance tests are to be conducted, then the absolute numbers do matter, but that is a separate issue, and one we do not discuss here.) Since, moreover, the sum of the proportion of those who score above t who turn out to be bad, π_B, and the proportion of those who score above t who turn out to be good, π_G, is 1, either one of these will be sufficient. We take π_B as our measure. That is, *the proportion of the applicants who score above t who eventually turn out to be bad is the appropriate measure of effectiveness of an application scorecard*. We call this measure the *bad rate amongst accepts*, and denote it by M_1.

This result has implications for current practice. In particular, it means that many of the popular measures in widespread use may not really be suitable as measures of performance of application scorecards. These include the Gini coefficient, the Kolmogorov–Smirnov (KS) statistic, the mean difference (t-statistic), and the information value (or divergence),[13,14] all of which use irrelevant information and fail to use information on the bad rate among accepts. We define these measures below, and illustrate, for the Gini coefficient, how it fails to measure the aspects of performance in which we are really interested. Similar shortcomings apply to the other measures.

Let $s(x)$ be the score for a person with descriptive characteristic vector x. For applicants with scores above t only, let B be the cumulative distribution function of these scores for the bad class, let G be the cumulative distribution function of these scores for the good class, and let b and g be the corresponding probability density or mass functions. Using this notation:

- the Gini coefficient is defined as

$$2\times \int B(s)g(s)\mathrm{d}s - 1 \tag{1}$$

- the KS statistic is defined as

$$\max_s |B(s) - G(s)| \tag{2}$$

- the mean difference statistic is defined as

$$\frac{\int sg(s)\mathrm{d}s - \int sb(s)\mathrm{d}s}{(\pi_G\{\int s^2 g(s)\mathrm{d}s - (\int sg(s)\mathrm{d}s)^2\} + \pi_B\{\int s^2 b(s)\mathrm{d}s - (\int sb(s)\mathrm{d}s)^2\})^{1/2}} \tag{3}$$

- the information value is defined as

$$\int (g(s) - b(s))\log(g(s)/b(s))\mathrm{d}s \tag{4}$$

The key thing to note about all of these definitions is that they are based on the distributions of score values of the good and bad customers. That is, these measures use information about the difference between the threshold t and the score, not merely the proportions of applicants who score above t who are good and bad. Only in the definition of the mean difference do the proportions π_G and π_B appear, and even here not in a way central to the definition but only as a weighting in calculating an average standard deviation of the two classes. In summary, all four of the measures defined above use information which is irrelevant to the performance of the scorecard and fail to use information which is critically relevant. This is not of mere theoretical interest. It means that incorrect conclusions can be drawn about scorecard performance.

It is easy to contrive examples showing how the Gini coefficient and other measures can be misleading. Consider for example, the following situation. Suppose that 10% of the applicants are bad, that the overall score distribution of the goods is normal with mean 0.5 and standard deviation 1.0, $s\sim N(0.5,\ 1.0)$, that the bads have an overall score distribution $s\sim N(-0.5,\ 1.0)$, and that the threshold t is 0. This yields an overall accept rate of 65.32%. If these values are thought to be unrealistic for real scorecards, then they can be adjusted by simple rescaling. The bad rate amongst accepts based on these values is 0.0472 and the Gini coefficient (based solely on the distributions of those scoring above t, of course) is 0.3224. Now suppose that another scorecard applied to this population, or the same scorecard applied to a new population in the future, yields scores which have the same good distribution (and that the population still contains 10% bads and the threshold is still $t=0$), but that the overall distribution of scores for the bads is now $s\sim N(-0.2,\ 0.5)$. That is, the mean bad score has increased and the standard deviation of the bad scores has decreased. Using these new values, the bad rate among accepts is 0.0525 and the Gini coefficient (again, of course, based solely on those scoring above t) is 0.6338. That is, the Gini coefficient (for which higher scores represent improved performance) shows an apparent improvement, having almost doubled, even though the bad rate has deteriorated (in fact it has increased by over 10% on its previous value).

The threshold t may be chosen in various ways. In the example above, we used a fixed value ($t=0$)—on the grounds that, when the instrument was constructed, a given score corresponded to a given risk of being bad. On the other hand, one might choose t so that a fixed proportion of the applicant population had scores above it (eg so that 80% of the population of applicants are accepted). Note that one will generally not want simply to choose the threshold which leads to minimum bad rate among accepts, since this is

achieved by accepting very few applicants–only those who score very highly indeed—and it is likely that one would want to accept more than the very few applicants accepted by this procedure. Now compare the initial situation above (which led to a Gini coefficient of 0.3224 and a bad rate amongst accepts of 0.0472) with a new situation which has arisen with a new population. For this new population, suppose that the score distribution of the goods is again $s \sim N(0.5, 1.0)$, and that the score distribution of the bads is $s \sim N(-0.25, 0.5)$. Only now assume that the overall proportion of bads in the new population is no longer 10% but is now 21.1%. With a threshold of $t = -0.1$ the accept rate is maintained at 65.32%. However, the Gini coefficient becomes 0.6448, again showing a dramatic increase on the original figure 0.3224 and again suggesting a substantial improvement, while the bad rate among accepts has increased to 0.1234. This is a three-fold increase on the previous bad rate, showing that in fact things have become substantially worse, so that the improved Gini coefficient is very misleading.

When neither action is rejection; for given threshold

The previous section considered the case when applicants with scores below the threshold t were not granted the product, so that they were never followed up and their true class was never discovered. For this reason, such customers could not appear in the scorecard's performance measure. In this section, we examine the case when the scorecard is used to assign customers to two different action classes, neither of them being rejection. In fact, for simplicity, we shall assume that those with scores above t will be regarded as 'reliable' customers, requiring no special action to be taken, while those with scores below t will be regarded as behaving in a risky manner, and requiring some kind of intervention. The details will depend on the aims, and the scorecard will be constructed to reflect those aims. In all these cases, in this section, we assume that the true class of all customers is eventually discovered, regardless of whether their score was above or below t. Again for convenience, we will refer to the classes as 'good' and 'bad', and take a high score as indicating that a customer is more likely to belong to the good class.

Once again, regardless of the extent to which their score exceeds t, the same (ie no) action will be taken for all customers whose score lies above t. Likewise, some other 'same action' will be taken for all customers whose score lies below t, regardless of the size of the difference. Thus only the counts of good customers above t, of good customers below t, of bad customers above t, and of bad customers below t provide information relevant to scorecard performance. Note that the counts of good and bad customers with scores falling below the threshold, and the counts of good and bad customers with scores falling above the threshold, are the counts of those customers who *would* turn out to be good or bad if we carried out the action appropriate for those *above* the threshold. Thus the performance measure is a measure of how effective the scorecard is in separating out those for whom we should take the action appropriate for scores above t from those for whom we should do something different. A straightforward extension of these ideas collects data on the outcomes of both actions for the entire range of scores, and combines the effectiveness of the score at separating goods from bads under both actions. However, for simplicity we will not discuss this here. In practice, of course, one would normally not have samples of customers with scores below the threshold who have received the action for those above the threshold, so that specific data collection strategies must be developed (eg give a random sample of customers one action, and another random sample the other action). The case in which one action is reject does not fit into this situation simply because the notion of a good or bad outcome when the customer is rejected has no meaning.

The definitions of the Gini coefficient, the KS statistic, the mean difference statistic, and the information value given in the previous section also apply here, except that now the distribution $g(s)$ is over the entire range of scores, and is the distribution of scores for customers who would turn out to be good if the action appropriate to scores above t were to be taken. Likewise for the bad distribution $b(s)$. Similarly, the proportion $\pi_G = 1 - \pi_B$ is the overall proportion good in the population under the action appropriate for those scoring above t. Thus, in this case as well, these measures all use (irrelevant) information about the difference between the scores and the threshold, while failing to use (critically) relevant information about the counts of customers lying on each side of the threshold.

Suppose that a marginal cost c_B is incurred for any customer scoring above t who is in fact bad, and that a marginal cost c_G is incurred for a customer scoring below t who turns out to be good. These costs arise from the fact that, for such people, an inappropriate action is recommended. The word *marginal* here refers to the fact that we are talking about the extra cost due to the fact that an *incorrect* prediction has been made, not the cost due to the effort of making a classification, which we assume to be the same in all cases.

The costs may be difficult to determine, but in fact there is a close relationship between them and the optimal choice of threshold t. To see this, recall that a score is really a monotonically increasing transformation of an estimate of the probability that a customer will belong to the good class: higher scores correspond to higher probabilities. We could recalibrate the score so that it directly yielded an estimate of this probability. Suitable (monotonic increasing) transformations can be estimated from the test data, as we describe below. Suppose that P is such a transformation, so that $P(s)$ is the estimated probability that a customer with score s will belong to the good class. Now, if $n_B(t)$ 'bad' customers have

estimated probability of being good larger than $P(t)$ and $m_G(t)$ 'good' customers have estimated probability below $P(t)$, the overall misclassification cost is $n_B(t)c_B + m_G(t)c_G$. It is then easy to show[15,16] that the optimal threshold, t, in the sense that it minimizes the overall cost, is given by $t = P^{-1}(c_B/(c_G + c_B))$. This analysis can be worked in reverse. If a threshold t is adopted, then the implicit ratio between the costs of the two types of misclassification is $c_G/c_B = (1-P(t))/P(t)$. The choice of threshold equal to t means that the scorecard users regard the misclassification of a good customer as $(1-P(t))/P(t)$ times as serious as the misclassification of a bad customer.

It follows from this that scorecards can be compared using the overall cost consequent on recommending inappropriate actions for those 'good' customers with scores below threshold t and those 'bad' customers with scores above t. The predictive performance of the scorecard is indicated by the overall cost due to misclassifications:

$$n_B(t)c_B + m_G(t)c_G = c_B\left(n_B(t) + m_G(t)\frac{1-P(t)}{P(t)}\right) \quad (5)$$

Since c_B, the cost of taking inappropriate action on a bad customer, is a constant for a fixed t, and since the total number, n, of customers to which the scorecard is applied should not influence the measures, we use, as our performance measure

$$M_2 = \left(n_B(t) + m_G(t)\frac{1-P(t)}{P(t)}\right)\Big/ n \quad (6)$$

Again we note the empirical data which goes into this measure. It is simply the counts of those customers who are assigned to inappropriate classes, so that inappropriate action is recommended for them. There is no use of irrelevant information about sizes of scores beyond the information relating to the action the scores imply. The larger is the value of M_2, the worse is the scorecard.

In order to use this measure, the value $P(t)$ needs to be determined. This can be carried out in various ways. Perhaps the most straightforward is to fit a simple logistic regression model with the response being the true class (with good labelled 1 and bad labelled 0) under action appropriate for those who score above t, and the predictor variable being the score on the scorecard. The value predicted from this logistic model at t gives an estimate of $P(t)$. If a logistic model is felt to be too restrictive, a straightforward alternative is to use local smoothing. For example, the local logistic model described in Tibshirani and Hastie.[17] More generally, logistic models (or, indeed, any model which yields a monotonic increasing relationship between score and estimated probability of being good) can also be used for recalibrating scorecards[18] so that, for example, after recalibration, a given score corresponds to a specified log odds of being good.

To illustrate the use of the measure given in (6), and to show how other measures can again lead to mistaken conclusions, consider the following situation. We illustrate using the KS statistic. Suppose that 10% of the applicants are bad, that the score distribution of the goods is normal with mean 1.0 and standard deviation 1.0, $N(1.0, 1.0)$, that the bads have a score distribution $N(-0.5, 1.0)$, and that the threshold t is 0. Then the KS statistic is 0.5467 and $M_2 = 0.1533$. Now suppose that this scorecard is applied at a later date to a population for which the distribution of scores of the goods becomes $N(2, 1.0)$, the distribution of scores of the bads becomes $N(0, 1.0)$, and the proportion of bads in the population is still 10% (and the threshold remains at 0). (Or imagine that an alternative scorecard applied to the original population yields distributions with these characteristics.) With these new parameters, the KS statistic is 0.6827 and $M_2 = 0.7722$. We see that whereas the KS statistic (for which a large value is good) shows an apparent large improvement in performance of the scorecard from the first to the second situation, the M_2 measure (for which a large value is bad) shows that there has been a dramatic deterioration in performance.

When neither action is rejection: unknown threshold

In the previous section, we described how to measure the overall costs associated with using a scorecard. However, this derivation was based on the assumption that the threshold t, which was used in assigning customers to classes, was known or given. Of course, in order to use the scorecard in practice one must choose a threshold—or, equivalently, choose a cost ratio c_G/c_B. However, at the time of building the scorecard and choosing between scorecards, one may not know exactly what a suitable threshold or cost ratio will be. The problem is that the precise future circumstances in which the scorecard will be applied are often unknown at the time the scorecard is selected. This would make formal performance criteria based on an assumption of a known threshold impossible to calculate at the time of constructing or selecting the scorecard.

The Gini coefficient can be regarded as being a way to tackle this. With threshold t, the overall cost is $C(t) = \{\pi_G G(t)c_G + \pi_B[1-B(t)]c_B\}$. If the scorecard classification threshold t is assumed to be unknown, we can obtain an overall measure by integrating $C(t)$ over all possible values of t. Of course, some values of t might be regarded as more likely than others. Let the probability that a value for t will occur be given by a function $w(t)$. For illustration, and because it will be seen to be an important special case, suppose we assume that $w(t) = \pi_G g(t) + \pi_B b(t)$, the mixture distribution of the scores. With this weight function, the

overall cost is

$$C = \int \{\pi_G G(t) c_G + \pi_B [1 - B(t)] c_B\} w(t) \mathrm{d}t \quad (7)$$

$$= \int \{\pi_G G(t) c_G + \pi_B [1 - B(t)] c_B\} [\pi_G g(t) + \pi_B b(t)] \mathrm{d}t \quad (8)$$

Without loss of generality, we can take $c_G + c_B = 1$. Then, using the relationship $c_G/c_B = (1-P(t))/P(t)$ and the fact that $P(t)$ is the probability that a customer with a score t will be good, we obtain

$$c_G = \pi_B b(t)/[\pi_G g(t) + \pi_B b(t)] \quad (9)$$

and

$$c_B = \pi_G g(t)/[\pi_G g(t) + \pi_B b(t)] \quad (10)$$

Substituting (9) and (10) into (8) yields

$$C = \int \{\pi_G G(t) \pi_B b(t) + \pi_B [1 - B(t)] \pi_G g(t)\} \mathrm{d}t \quad (11)$$

from which

$$C = \pi_G \pi_B \int \{G(t) b(t) + [1 - B(t)] g(t)\} \mathrm{d}t$$
$$= 2\pi_G \pi_B \left[1 - \int B(t) g(t) \mathrm{d}t\right]$$

which, from (1), we see gives

$$C = \pi_G \pi_B (1 - Gini) \quad (12)$$

and

$$Gini = 1 - C/\pi_G \pi_B \quad (13)$$

Thus, if we do not know what threshold to choose, and if we combine the costs over *all possible* thresholds, where the choice of threshold is weighted according to the overall score mixture distribution, we obtain a result which is equivalent to (a simple linear transformation of) the Gini coefficient.

The choice of the mixture distribution as the weighting function $w(t)$ in the derivation of (12) and (13) may appear rather artificial. Indeed, it is. The costs c_B and c_G will in fact be obtained quite distinctly from the score distribution, and need have no relationship to it whatsoever. They should be determined by the consequences of the actions, and the context in which those actions occur (the type of financial product, etc). Furthermore, the integral in the above derivation ranges over all possible values of the score. This is equivalent to ranging over all possible values of the cost ratio c_G/c_B, from 0 to ∞. It means, for example, that one is prepared to include in the measure the possibility that the cost of misclassifying a good customer is 10 times as serious as misclassifying a bad customer, *and* the possibility that the cost of misclassifying a good customer is only one tenth as serious as misclassifying a bad customer. This will seldom be appropriate. Usually, one will believe that one type of cost will be larger than the other. For example, treating a bad customer as good often leads to financial loss whereas treating a good customer as bad often only means a lost opportunity. If it is known that one type of cost is larger than the other, then the cost ratio should range either over the interval [0, 1] or over the interval [1, ∞]. More generally, one might be prepared to give likely ranges for the cost ratio or threshold. We return to this below.

In summary, the advantage of the Gini coefficient is that it requires no thought about the possible range of the threshold or, equivalently, of what the cost ratio c_G/c_B might be. However, the converse of this advantage is the major disadvantage that it removes the need to think about the range of the cost ratio by integrating over *all possible* values—even though we know that some of these values are entirely unrealistic for any given product—and gives relative weights to the possible values, via $w(t)$, which may bear no relationship whatsoever to the likely values for the threshold.

The relationship $c_G/c_B = (1-P(t))/P(t)$, with $c_G + c_B = 1$, gives $c_B = P(t)$. Using this in (7) gives

$$C = \int \{\pi_G G(t)[1 - P(t)] + \pi_B [1 - B(t)] P(t)\} w(t) \mathrm{d}t \quad (14)$$

Now, defining c_G as $\int_{-\infty}^{\infty} G(t)[1 - P(t)] w(t) \mathrm{d}t$, we have

$$C_G = \int_{-\infty}^{\infty} G(t)[1 - P(t)] w(t) \mathrm{d}t$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{t} g(u) \mathrm{d}u [1 - P(t)] w(t) \mathrm{d}t$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{t} g(u)[1 - P(t)] w(t) \mathrm{d}u \mathrm{d}t$$
$$= \int_{-\infty}^{\infty} \int_{u}^{\infty} g(u)[1 - P(t)] w(t) \mathrm{d}t \mathrm{d}u$$
$$= \int_{-\infty}^{\infty} g(u) \int_{u}^{\infty} [1 - P(t)] w(t) \mathrm{d}t \mathrm{d}u$$
$$= E_G \left(\int_{u}^{\infty} [1 - P(t)] w(t) \mathrm{d}t \right)$$

where the notation E_G signifies that the expectation is taken with respect to the distribution G.

We can estimate this expectation using

$$\hat{C}_G = \frac{1}{n_G}\sum_{i=1}^{n_G}\int_{t_i}^{\infty}[1-P(t)]w(t)\mathrm{d}t \tag{15}$$

where the t_i, $i=1,\ldots,n_G$ are the scores of the sample points from the good class. This can be further expressed as

$$\hat{C}_G = \frac{1}{n_G}\sum_{i=1}^{n_G} i\int_{t_i}^{t_{i+1}}[1-P(t)]w(t)\mathrm{d}t \tag{16}$$

where $t_{n_G+1}=\infty$. This choice of t_{n_G+1} might suggest that the last integral in this summation will be infinite. However, this will only be the case if one is prepared to contemplate values which are larger than the observed scores of all the good customers (so that they are all classified as bad). It seems unlikely that one would ever be prepared to contemplate such a value for t. This restriction on possible values for t will be made explicit in the choice of $w(t)$, which will be zero for such large values of t.

A similar calculation for the second term in (14) yields

$$C_B = \int_{-\infty}^{\infty}[1-B(t)]P(t)w(t) = E_B\left(\int_{-\infty}^{u}P(t)w(t)\mathrm{d}t\right)$$

which may be estimated by

$$\hat{C}_B = \frac{1}{n_B}\sum_{j=1}^{n_B}(n_B-j+1)\int_{t_{j-1}}^{t_j}P(t)w(t)\mathrm{d}t \tag{17}$$

where the t_j, $j=1,\ldots,n_B$ are the sample points from the bad class and $t_0=-\infty$. An argument similar to that for the good class means that $w(t)$ will be zero for values of t smaller than the smallest observed score for bad customers.

Adding (16) and (17), appropriately weighted by the priors π_G and π_B respectively, gives an estimate of the total cost (14).

Since $P(t)$ and $w(t)$ will generally be fairly smooth functions, and will not change much between two customers with neighbouring scores, we can approximate (14) by

$$\begin{aligned}\tilde{C} = {}& \frac{\pi_G}{n_G}\sum_{i=1}^{n_G} i(t_{i+1}-t_i)\left(1-\frac{P(t_{i+1})+P(t_i)}{2}\right)\\ &\times\left(\frac{w(t_{i+1})+w(t_i)}{2}\right)+\frac{\pi_B}{n_B}\sum_{j=1}^{n_B}(n_B-j+1)(t_j-t_{j-1})\\ &\times\left(1-\frac{P(t_j)+P(t_{j-1})}{2}\right)\left(\frac{w(t_j)+w(t_{j-1})}{2}\right)\end{aligned} \tag{18}$$

If the data have been obtained by a random sample from the population, so that reasonable estimates of π_G and π_B are, respectively, $n_G/(n_G+n_B)$ and $n_B/(n_G+n_B)$, (18) becomes

$$\begin{aligned}\tilde{C} = {}& \frac{1}{n_G+n_B}\left\{\sum_{i=1}^{n_G} i(t_{i+1}-t_i)\left(1-\frac{P(t_{i+1})+P(t_i)}{2}\right)\right.\\ &\times\left(\frac{w(t_{i+1})+w(t_i)}{2}\right)+\sum_{j=1}^{n_B}(n_B-j+1)(t_j-t_{j-1})\\ &\left.\times\left(1-\frac{P(t_j)+P(t_{j-1})}{2}\right)\left(\frac{w(t_j)+w(t_{j-1})}{2}\right)\right\}\end{aligned} \tag{19}$$

The derivation of (18) is perfectly general, and applies for any choice of $w(t)$ (subject to the constraints that it takes zero values for very extreme scores, so that the integrals and sums above are not infinite). Formal methods of knowledge elicitation could be used to try to extract beliefs about likely values of t, that is, about the function $w(t)$. However, it seems unlikely that the effort involved, and the accuracy which will result, will justify the effort. Instead, therefore, we propose a simple approximate method, following that described in Adams and Hand.[19] This is based on obtaining just three possible values for t: the value thought to be most likely (t_M), the value which is regarded as the lower limit of possible values (t_L), and the value which is regarded as the upper limit of possible values (t_U). Using these values, a simple triangular form is defined for $w(t)$:

$$w(t) = \begin{cases} 0 & t<t_L \\ \dfrac{(t-t_L)}{2(t_M-t_L)(t_U-t_L)} & t_L<t<t_M \\ \dfrac{(t_U-t)}{2(t_U-t_M)(t_U-t_L)} & t_M<t<t_U \\ 0 & t>t_U \end{cases} \tag{20}$$

This places most weight in the region of t_M and decays linearly to zero weight at t_L and t_U. It is scaled to integrate to 1.

Some other important issues

In the introduction, we mentioned some general practical issues, which affected all scorecard construction methods, and which needed to be taken into account when assessing scorecards. There are also other high-level issues which should be kept in mind, both when building and when evaluating scorecards. Without going into too much detail, we will describe some of the most important here.

There is an important distinction between evaluating scorecard construction methods and evaluating scorecards. Thus, for example, several authors[20,21] have undertaken comparative studies of the effectiveness of neural networks, support vector machines, and other recent developments

compared with logistic regression, linear discriminant analysis, and tree methods. Their aim, in these studies, is to draw some general conclusion about which class of methods is most effective (for retail banking applications). This will help those tasked with constructing new scorecards to choose between methods. In contrast to this, there is the situation of being presented with a data set and invited to construct various scorecards and choose the best for practical application. The aim here is to draw a conclusion about which scorecard is best *with this data set*. The problems have been termed the *unconditional* and *conditional* problems, respectively. The conditional one makes a statement specific to (conditional on) the available data, while the unconditional one is more general. It is entirely possible that opposite conclusions can be drawn from the two questions: for example, perhaps in general neural networks provide good solutions, but it may be that, in the particular case of the data presented to us, the network solution is poor. It is therefore very important to be clear about which of the questions one is trying to answer.

When one assesses performance over time to detect deterioration, new data are constantly arriving. The scorecard is being applied to the data set on which it is to be evaluated (subject to difficulties of the kind mentioned at the end of the introduction). However, when one is initially constructing a scorecard or choosing which of several possible scorecards to adopt, one only has available the construction data. It is well known[14,16,22] that evaluating a scorecard on the data used to construct it (ie, deciding which characteristics to include, choosing how to split the characteristics into categories, estimating parameters and weights, and so on) leads to optimistic evaluations, in the sense that future performance is likely to be worse than that estimated and some highly sophisticated strategies have been developed for overcoming it.[14,16,22]

This paper is concerned with evaluating scorecards. Such evaluation might be for absolute or comparative purposes. For example, in application scoring we might want to be sure that the bad rate among accepts is kept below 5%—an absolute value of performance. Or we might want to know which of two scorecards is more effective—a comparative evaluation. Again it is important to bear the distinction in mind.

Finally, scorecards are often constructed using statistical model-fitting approaches, optimising the criteria typically adopted in such modelling, such as likelihood. Other criteria include least squares with neural networks, various impurity indices with tree classifiers, and distance from the decision surface with support vector machines. Since different criteria can lead to different results, it would seem sensible to choose the scorecard by optimizing a criterion appropriate to the use to which the scorecard is to be put—such as those described in this paper.

Conclusion

If a scorecard is being used to assign customers to actions by comparing their score with a threshold, and if the same action will be taken and penalty will be incurred no matter how large is the difference between the score and the threshold, then scorecard performance criteria should depend only on the numbers of customers assigned to the actions, and not on how close the scores are to the decision threshold. If outcomes will result from both possible actions, then measure M_2, defined in Equation (6), is appropriate. A complication arises in cases when one of the actions is to reject an applicant, since then by definition the outcome does not exist, so that it is never known what would have been the appropriate action to choose. In this case, performance criteria must be based solely on the accepted applicants, and M_1, the bad rate among accepts is appropriate.

It is important to note that we are not making a blanket statement to the effect that the measures in common use, such as the Gini coefficient, the KS statistic, the mean difference, and the information value, are never appropriate. Rather, we are making (a) the broad statement that the choice of measure must reflect the aims of the scorecard procedure, and (b) the narrower statement that, in the case when the choice of action depends solely on whether a person scores above or below some threshold, these common measures use irrelevant information, which means that they may draw misleading, and even incorrect conclusions.

A final cautionary note is in order. We have ignored the subtleties raised at the end of the introduction. These are important in practical implementations, but affect all scorecard applications and assessments, not merely the criteria we have described in this paper.

Acknowledgements—The work in this paper was stimulated by discussions with Mark Kelly, Ian Warren, and others from Fair, Isaac, and was partially supported by Fair, Isaac. I am especially grateful to Lyn Thomas for his helpful suggestions on an earlier draft of this paper, and to all those who commented after I presented some of the material at the *Credit Scoring and Credit Control VIII* conference in Edinburgh, September 2003, and at the *Credit Rating and Scoring Models Conference* in Alexandria, May 2004.

References

1 Rosenberg E and Gleit A (1994). Quantitative methods in credit management: a survey. *Opns Res* **42**: 589–613.
2 Hand DJ and Henley WE (1997). Statistical classification methods in consumer credit scoring: a review. *J R Stat Soc Ser A* **160**: 523–541.
3 Hand DJ (1998). Consumer credit and statistics. In: Hand DJ and Jacka SD (eds). *Statistics in Finance*. Arnold, London, pp 69–81.
4 Hand DJ (2001a). Modelling consumer credit risk. *IMA J Mngt Math* **12**: 139–155.

5 Thomas LC (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *Int J Forecasting* **16**(2): 149–172.
6 Thomas LC, Edelman DB and Crook JN (2002). *Credit Scoring and its Applications*. SIAM: Philadelphia.
7 Hand DJ and Henley WE (1993). Can reject inference ever work? *IMA J Math Appl Business Ind* **5**: 45–55.
8 Hand DJ (2001b). Reject inference in credit operations. In: Mays E (ed). *Handbook of Credit Scoring*. Glenlake Publishing, Chicago, pp 225–240.
9 Crook J, Banasik J and Thomas L (2001). Sample selection bias in credit scoring. *Credit Scoring and Credit Control VII*, Management School, University of Edinburgh.
10 Kelly MG, Hand DJ and Adams NM (1999). The impact of changing populations on classifier performance. In: Chaudhuri S and Madigan D (eds). *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, USA, pp 367–371.
11 Hand DJ and Kelly MG (2001). Lookahead scorecards for new fixed term credit products. *J Opl Res Soc* **52**: 989–996.
12 Hand DJ (2001c). Measuring customer quality. *Credit Scoring and Credit Control VII*, Management School, University of Edinburgh.
13 Wilkie AD (1992). Measures for comparing scorecard systems. In: Thomas LC, Crook JN, and Edelman DB (eds). *Credit Scoring and Credit Control*. Clarendon Press: Oxford, pp 123–138.
14 Hand DJ (1997). *Construction and Assessment of Classification Rules*. Wiley: Chichester.
15 Hand DJ (1981). *Discrimination and Classification*. Wiley: Chichester.
16 Webb AR (1999). *Statistical Pattern Recognition*. Arnold: London.
17 Tibshirani R and Hastie T (1987). Local likelihood estimation. *J Am Stat Assoc* **82**: 559–567.
18 Thomas LC, Banasik J and Crook JN (2001). Recalibrating scorecards. *J Opl Res Soc* **52**: 981–988.
19 Adams NM and Hand DJ (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recogn* **32**: 1139–1147.
20 Davis RH, Edelman DB and Gammerman AJ (1992). Machine learning algorithms for credit card applications. *IMA J Math Appl Business Ind* **4**: 43–51.
21 West D (2000). Neural network credit scoring models. *Comput Opns Res* **27**: 1131–1152.
22 Ripley BD (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press: Cambridge.

Received October 2003;
accepted October 2004 after one revision