Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Banking & Finance 28 (2004) 857–874
Journal of BANKING & FINANCE
www.elsevier.com/locate/econbase

Does reject inference really improve the performance of application scoring models?

Jonathan Crook *, John Banasik [1]

Credit Research Centre, The School of Management, WRB, University of Edinburgh, 50 George Square, Edinburgh, Scotland EH8 9JY, UK

Abstract

The parameters of application scorecards are usually estimated using a sample that excludes rejected applicants which may prove biased when applied to all applicants. This paper uses a rare sample that includes those who would normally be rejected to examine the extent to which (1) the exclusion of rejected applicants undermines the predictive performance of a scorecard based only on accepted applicants, and (2) reject inference techniques can remedy the influence of this exclusion.

JEL classification: C24; C44; C51; D14
Keywords: Risk management; Credit scoring; Reject inference; Augmentation; Extrapolation

1. Introduction

Application credit scoring is the process of predicting the probability that an applicant for a credit product will fail to repay the loan in an agreed manner. To assess this process we require a model that represents the behavior of all applicants for credit, yet typically we have only information about the repayment behavior of those who have been accepted (and booked) for credit in the past. The behavior of those who have been rejected, if instead they had been accepted, is unknown. If one estimates a model using data only on accepted applicants, those estimated parameters may be biased when applied to all applicants. In addition, if cut-off scores are chosen to equalize the actual and predicted number of defaulting applicants then a sample of

* Corresponding author. Tel.: +44-131-650-3802; fax: +44-131-668-3053.
E-mail addresses: j.crook@ed.ac.uk (J. Crook), john.banasik@ed.ac.uk (J. Banasik).
[1] Tel.: +44-131-650-3793.

0378-4266/$ - see front matter
doi:10.1016/j.jbankfin.2003.10.010

accepted applicants is likely to yield inappropriate cut-offs for the population of all applicants.

Reject inference techniques attempt to incorporate characteristics of rejected applicants into the process of calibrating a scorecard based primarily on the repayment behavior of accepted applicants. Various reject inference techniques have been proposed either in the literature or by consultancies. [2] Relatively little has been published that empirically compares the predictive performance of algorithms that incorporate different possible reject inference techniques. Meester (2000), Banasik et al. (2003), and Ash and Meester (2002) are examples. There is no published empirical evaluation of the predictive performance of the reject inference technique that is perhaps the most frequently used, augmentation (or re-weighting). The aim of this paper is to report such an evaluation and to compare its performance with another reject inference technique, extrapolation. These results may in turn be compared very precisely with those obtained by Banasik et al. (2003) using bivariate probit on the same samples. Re-weighting and bivariate probit are superficially similar techniques in some respect, but they typically depend on different premises and accomplish distinct outcomes.

This paper analyses a rare sample drawn by a credit-provider who occasionally grants credit to virtually all applicants in order to avoid biased estimation of scorecard parameters. Availability of normally rejected applicants in the data set permits evaluation of how their absence from the data set would undermine prediction performance. Scorecards calibrated with and without these applicants can be applied to all applicants in order to assess the extent of inaccuracy in the normal situation where they are excluded from scorecard modelling. After assessing the extent of inaccuracy arising from absence of repayment behavior of rejected applicants, the available data sets permits consideration of the extent of corrective influence achieved by some reject inference techniques. In particular, this paper considers the efficacy of two reject inference techniques, for five acceptance thresholds and for two commonly used types of predictor variable.

Re-weighting, extrapolation, and bivariate probit approaches are all techniques that address a sample selection problem. Non-random selection of applicants results in applicants with some characteristics being disproportionately present in the sample or perhaps not present at all. Heckman (2001) characterizes this type of problem as a "weighting" problem in that various types of applicants have inappropriate weights or even zero weights applied to them in the process of sampling. However, re-weighting as the term is conventionally used may not fix this problem, since the re-weighting of selected cases cannot retrieve types of applicants that were zero-

[2] These include extrapolation, augmentation (Hsai, 1978), iterative reclassification (Joanes, 1993/4), bivariate probit (Boyes et al., 1989), "parcelling", use of the EM algorithm (Demster et al., 1977), using a multinomial logistic model (Reichert et al., 1983), and collecting repayment performance data for rejects (Hand and Henley, 1993/4; Ash and Meester, 2002). The plausibility of the necessary assumptions in using these techniques with data typically used in credit scoring models has been reviewed by a number of authors (Ash and Meester, 2002; Banasik et al., 2003; Hand and Henley, 1993/4; Joanes, 1993/4; Thomas et al., 2002).

weighted in the selection process. Re-weighting can produce biased weighting for those types of cases that are represented.

Extrapolation, the other reject inference technique examined in this paper, imputes a good–bad classification to rejected cases on the basis of an initial model estimated using only accepted applicants. A final model then can be estimated using all applicants. Perhaps the main motivation for re-weighting and for extrapolation is a desire to accommodate the possibility that a single set of parameters does not govern all applicants to be scored by the good–bad model. The most appropriate parameter estimates to adopt would be an "average" of parameter values pertaining to accepted and rejected applicants. This would tend to focus attention most on discriminating among applicants whose apparent creditworthiness is most marginal. In the absence of observed repayment performance for rejected applicants "average" values might be best approximated by giving greater weight to accepted applicants who resemble those who were rejected.

In contrast to this, bivariate probit with sample selection presumes that all cases are governed by a single model distinguishing good applicants from bad. Its main concern is the influence of missing variables arising from exclusion of rejected applicants. For example, were all unemployed applicants to have been rejected, employment status would be a variable unavailable for subsequent modelling of repayment behavior. Moreover, acceptance may have been on the basis of variables that are no longer available, no longer suitable, or simply unknown in nature. A special instance of such missing variables arises from the common practice of loan officers overriding the result of the acceptance model, implying the deployment of notional additional variables. Bivariate probit attempts to retrieve the influence of missing variables that are reflected in the prior acceptance decision. The re-weighting method also resorts to an attempt to estimate the previous model, but does so with a view to restore the influence of missing cases by altering the relative weight allocated among the accepted cases. The previous model may have been distinct from that currently contemplated, but that is an incidental feature of the re-weighting approach.

In the next three sections general features of a credit scoring model are followed by more detailed descriptions of what re-weighting and extrapolation involve. In sections that follow we explain our methodology and results. The final section concludes.

2. A primer on credit scoring for economists

Logistic regression is a simple and appropriate technique for estimating the log of the odds of default as a linear function of application attributes:

$$\ln\left[\frac{\hat{P}(\text{Default})}{1-\hat{P}(\text{Default})}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k. \tag{1}$$

Typically the explanatory variables may be expressed, alternatively, as a set of dummy variable regressors, or as single regressors measuring the weights of evidence.

Table 1
Repayment behavior for ranges of years on electoral roll at current address

Complete years at electoral address	Under 1 year	1 year	2–3 years	4–7 years	8–10 years	Not known	Total cases
Bads (defaulters)	1333	165	178	168	204	105	2153
Goods (repayers)	1744	353	577	640	838	141	4293
Bads/goods	0.7643	0.4674	0.3085	0.2625	0.2434	0.7447	0.5015
Relative odds	1.5241	0.9320	0.6151	0.5234	0.4854	1.4849	1.000
Weights of evidence	0.4214	−0.0704	−0.4859	−0.6474	−0.7228	0.3953	

Consider Table 1. The explanatory variable, years on electoral roll at current address, has been *coarse classified* into a relatively small number of categories, primarily with a view to simplification and parsimony in the number of variables. It can then be represented by a set of five dummy variables, the first equal to 1 if "under 1 year" and zero otherwise, the second equal to 1 if "1 year" and zero otherwise, and so on. The "not known" category has arbitrarily been chosen as the omitted variable. Adoption of this set of regressors in the equation permits a simplified and controlled non-linear relationship to be expressed.

Alternatively, the years on electoral roll at current address variable can be expressed as a single regressor in which all cases in each class are attributed a single value called the *weight of evidence*. The values assigned to each class reflect the nature of the logistic regression that predicts the logarithm of the ratio $P(\text{Bad})/P(\text{Good})$ where bad means defaulted and good means non-defaulter. The weight of evidence assigned to each class is the logarithm of the $P(\text{Bad})/P(\text{Good})$ for that class less the logarithm of the same ratio that applies to the whole sample. For example in the first class the weight of evidence would be $\ln(0.7643/0.5015) = \ln(1.5241) = 0.4214$ which is the value assigned to every case in that class. If years on the electoral roll were the only variable, this weights of evidence variable would fetch a coefficient of unity and this single-regressor model would be equivalent to the five regressor model involving dummy variables.

Once estimated, using a portion of the sample (a training sample), the logistic regression model provides an estimated probability of default for each case in both the training and the hold-out sample (i.e. the remaining observations). In predicting defaulters these probabilities are generally taken as only *relative* indicators of a propensity to default. Instead of focusing on cases with a predicted probability above 0.5, one observes the estimated probability above which there are as many cases as there are observed defaulters. In order to remove the influence of over-fitting from assessment of model performance, both the estimated parameters of the model and that probability cut-off are estimated on a training sample and applied to a hold-out sample. Table 2 cross-tabulates actual and predicted good and bad cases.

The ROC (Receiver Operating Characteristic) curve provides a more general performance measure that avoids the influence of an arbitrarily chosen probability cut-off point. Consider the applicants ordered from worst to best in terms of their estimated probability of default. Two functions of this score are plotted against each

Table 2
Performance for a multivariate weights of evidence model for English cases

Training sample English cases					*Hold-out sample English cases*				
Actual	Predicted			Percent correct	Actual	Predicted			Percent correct
	Good	Bad	Total			Good	Bad	Total	
Good	3432	861	4293	79.94	Good	1705	441	2146	79.45
Bad	861	1292	2153	60.01	Bad	408	668	1076	62.08
Total	4293	2153	6446	73.29	Total	2113	1109	3222	73.65

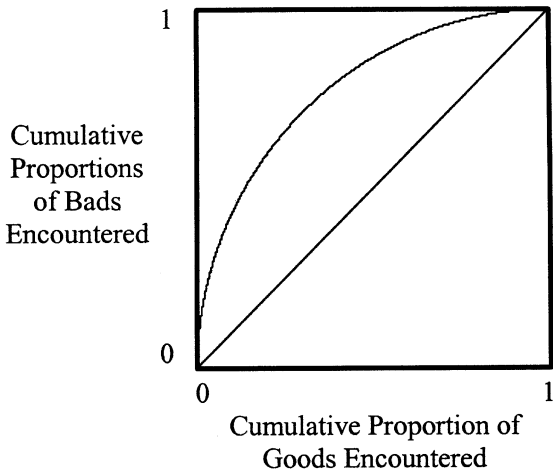

Fig. 1. ROC curve.

other, the cumulative proportion of bads and of goods. The curve illustrated in Fig. 1 indicates the success of the scoring model in distinguishing bad applicants from good. The straight diagonal line represents the performance of random selection whereby one draws bads and goods in proportion to their preponderance in the sample. The higher the curve above this diagonal line, the better, and overall performance is measured as the Area Under the ROC curve (AUROC).

3. Re-weighting

Although there are several variants of re-weighting, the basic method is as follows (see Table 3). First an accept–reject (AR) model is estimated using cases which have been accepted or rejected over a given period of time by the current model. If the model has been applied without over-rides, if the explanatory variables within it are known (call them X_{old}), and if the algorithm used, the functional form of the model, and all other parameters of the original estimation process are known, then this model can be estimated perfectly. Otherwise it cannot be estimated perfectly.

Table 3
Re-weighting

Band (j)	Number of goods	Number of bads	Number of accepts	Number of rejects	Band weight
1	g_1	b_1	$A_1 = g_1 + b_1$	R_1	$(R_1 + A_1)/A_1$
2	g_2	b_2	$A_2 = g_2 + b_2$	R_2	$(R_2 + A_2)/A_2$
–	–	–	–	–	–
–	–	–	–	–	–
–	–	–	–	–	–
–	–	–	–	–	–
–	–	–	–	–	–
n	g_n	b_n	$A_n = g_n + b_n$	R_n	$(R_n + A_n)/A_n$

Suppose we use a set of explanatory variables, X_{new}, to model the original accept–reject decision process. The scores predicted by this AR model for each case, $S_i = s(X_{\text{new},i})$, are divided into bands and within each band, j, the numbers of rejected R_j and accepted A_j cases are found. For each A_j there are g_j good cases and b_j bad cases.

Assume the bands are chosen so that within any band, j, the probability that an accepted applicant with score S_j is good equals the probability that a rejected applicant with score S_j is good, i.e.,

$$P(g|S_j, A) = P(g|S_j, R). \tag{2}$$

Then

$$g_j/A_j = g_j^r/R_j,$$

where g_j^r is the imputed number of goods amongst the rejects within band j; g_j^r/R_j is the proportion of rejects in band j that would have been good had they been accepted. The re-weighting technique weights the A_j accepts in band j to represent both the A_j and R_j cases in the band, i.e. each A_j is weighted by $(R_j + A_j)/A_j$. This is the inverse of the probability of acceptance in band j and is the probability sampling weight for band j.

Since accepted scores are monotonically related to the probability of being accepted we can replace scores by these probabilities, and if instead of bands we consider individual values, where there are m possible values (because there are m cases), each row in the expanded Table 3 relates to $P(A_i),\ i = 1 \ldots m$. Thus each accepted case has a probability sampling weight of $1/P(A_i)$. A good–bad model using the weighted accepts is then estimated.

In a world where a single set of parameters governs accepted and rejected applicants alike, one might presume that, although re-weighting would have no scope to eliminate bias, neither would it introduce bias. This is not necessarily so. If X_{old} is not a subset of X_{new} the model's parameter estimates will be biased. [3]

[3] See Little and Rubin (1987), Hand and Henley (1993/4, 1994), and Banasik et al. (2003).

In our initial analysis – hereafter referred to as investigation one – re-weighting involves adoption of the credit-provider's accept–reject status for all cases, and this was certainly based on a model including variables unknown to us. In the subsequent analysis – hereafter referred to as investigation two – the variable selection of the accept–rejection model has been contrived so as to differ from that of the good–bad model such that both models have variables excluded from the other. In short we simulate a world where X_{old} and X_{new} differ yet the modellers assume that they do not. We adopt this approach primarily to make the analysis realistic, particularly given the inevitable divergence between the credit supplier's model that governed its accept–reject decision and any subsequent model likely to be developed. Moreover, we recognize that this is probably the usual circumstance, particularly since the process of over-riding statistical models in the acceptance decision involves implicit introduction of other variables.

4. Extrapolation

As with re-weighting there are several methods of extrapolation. The method we consider is to estimate a posterior probability model using accept-only data, extrapolate the probability of default for the rejected cases and by applying a cut-off probability classify the rejected cases as either good or bad. A new good–bad model is then estimated for all cases (see Ash and Meester, 2002).

If the regression coefficients of the good–bad model that are applicable to the accepts also apply to the rejects then this procedure would have minimal effect on the estimates of these coefficients, although the standard errors of the estimated coefficients will be understated. However, if variables other than X_{new} affect the probability of acceptance, extrapolation would yield biased estimates of the posterior probabilities.

If X_{old} is a subset of X_{new} a further source of error in the predicted probabilities may occur due to the proportion of goods and bads in the training sample not being equal to the proportion in the all-applicant population. This may cause the cut-off probabilities, which equalize the expected and actual number of bads in the training sample, to deviate from the cut-offs required to equalize the actual and predicted number of bads in the all-applicant population. The regression model may give unbiased posterior probabilities, but applicants would be misallocated because inappropriate cut-offs may be applied.

5. Methodology

The proprietary nature of the available data set restricts the detail that can be described here, but its main characteristics can be set out. The data pertain to applicants for credit within a fixed period during 1997. To obtain credit a customer needed to progress through two stages. First, an applicant sought information about the product. Some potential applicants were rejected at this stage. No information about these applicants was provided, but the credit grantor has indicated that this

was a negligibly small proportion of applicants. Subsequent tables make plain the presence of a very high proportion of dubious applicants, and details of many individual cases confirm the presence of cases with appalling credit histories. Second, those who receive information apply for the product.

A repayment performance is defined to be "bad" if the account was transferred for debt recovery within 12 months of the credit being first taken. All other accounts were defined to be "good". We had available the accept–reject decision that the credit grantor would have implemented for each applicant under normal practice, although for our sample the indicated decision had not been implemented. This decision was deterministic – there were no overrides – and was based on an existing statistical model that had been parameterized from an earlier sample. Neither the existing model nor its sample were available to us, but the provided data set contains almost all of the variables needed to estimate the existing model. Only a relatively small subset of these provided variables was actually used to build that existing model. Little was indicated about the existing model beyond that.

In an earlier paper (Banasik et al., 2003) we indicated, using the same data set, that there was limited scope for reject inference to achieve an increase in predictive performance using the data supplier's acceptance threshold. Investigation one in this paper explores how much of that limited scope is achieved by re-weighting, again using the data supplier's acceptance threshold. As in the earlier paper we recognize that perhaps the main influence restricting the scope for improved predictive performance is the already very low acceptance threshold of the credit provider who normally accepts roughly two thirds of applicants of whom nearly a quarter are "bad". Such a sample already focuses ample attention on poor credit risks. Accordingly, investigation two explores how the scope for improved predictive performance and its achievement by reject inference might both be varied by alteration of the acceptance threshold. Analysis distinguishes two influences of lowering the acceptance threshold, achievement of improved ranking among cases and discernment of the cutoff score that will reflect the proportion of bads in the population.

Investigation one, which deals with the credit-grantor's own normal acceptance threshold, was based upon stratified proportional random samples that ensured the same good–bad ratio prevailed in training samples as in corresponding hold-out samples. Analysis involves two model comparisons. First, a model parameterized on a training sample that includes both accepted and rejected cases is contrasted with another parameterized on a training sample of only accepted cases. This establishes the scope for reject inference to improve prediction. Secondly, the model parameterized only on accepted cases is contrasted with one that benefits from some form of reject inference.

Investigation two, which considers the scope and achievement of reject inference under a variety of acceptance thresholds, is essentially the same as for investigation one. Unfortunately, since neither the data provider's existing model nor even the scores it generated are available, this analysis must depend on an initial accept–reject model fabricated to resemble that of the data provider yet is distinct from our own good–bad model. In this way it should reflect a normal situation in which a good–bad model becomes stale in terms of its selection of predictor variables and param-

eter estimates and is replaced by another. The initial model then becomes the accept–reject model that has determined (in conjunction with over-rides) which applicants are used for estimating the new good–bad model. [4]

The results of various experiments reported in the tables focus mainly on orders of magnitude and patterns thereof. When proportions are roughly 70% and apply to a pair of hold-out samples of 4069, as in investigation one, a standard error is roughly 1% and one-tail significance requires a percentage difference of 1.645%. Investigation two deals with the same sort of proportion and hold-out samples of 3222, so that such significance requires a difference of 1.878%.

6. Potential gains from re-weighting with original acceptance criterion

Our results when we used the data granter's classification of cases into accepts and rejects are shown in Tables 4 and 5. Tables 4 and 5 show the predicted performance using weights of evidence with and without re-weighting respectively. We used all of the 46 variables that were available to us.

Using area under ROC as the measure of accuracy (Comparison 1 in Tables 4 and 5) four observations can be made. (1) The scope for improvement by any reject inference technique is very small. Estimating an unweighted model for *accepted* applicants only (Table 4) and testing this on a hold-out sample of all applicants to indicate its true predicted performance gives an AUROC of 0.7818 compared to 0.7837 for a model estimated for a sample of *all* applicants. (2) Establishing an unweighted accept-only model and testing it on an accept-only hold-out sample overestimated the performance of the model. An accept-only model tested on an accept-only hold-out gave an AUROC of 0.7932 whereas the performance of the model tested on a sample of all applicants, including accepts and rejects, is 0.7818. (3) Using re-weighting as a method of reject inference was found to *reduce* the predictive performance of the model compared with an accept-only model; the AUROC values were 0.7765 (Table 5 Comparison 1) and 0.7818 (Table 4 Comparison 1) respectively. (4) Estimating a re-weighted model and testing it on an accept-only sample also overestimated the true performance, giving an AUROC of 0.7875 rather than a more representative 0.7765 (Table 5 Comparison 1).

The predictive performances using percentages correctly classified are also shown in Tables 4 and 5. Four observations can be made from these. (1) The scope for improvement due to improved model coefficients is small, from 71.74% to 72.13% (Table 4 Comparison 2). (2) The accept-only model tested on an accept-only hold-out sample (with training sample cut-offs) would considerably overestimate the model's performance: 76.19% correctly classified compared with 70.83% when tested

[4] A single coarse classification governs all models estimated in this paper. However, weights of evidence for all models have been estimated on the relevant training sample. All analysis was conducted using alternatively weights of evidence variables and dummy variables in order to test sensitivity of reject inference to this feature. Given the similarity of approaches using both types of variables, generally results for only weights of evidence are tabulated here.

Table 4
Original data: Simple logistic model with weights of evidence

Comparison 1: Area under ROC							
Predicting model	Training sample cases	Own band hold-out		All-applicant hold-out			Accept analysis delusion
		Number of cases	ROC area	Number of cases	ROC area		
Accepted	5413	2755	**0.7932**	4069	**0.7818**		**0.0114**
All case	8139	4069	**0.7837**	4069	**0.7837**		
Comparison 2: Percentage correctly classified							
Predicting model	Own band hold-out prediction			All-applicant hold-out prediction			Accept analysis delusion
	Number of cases	Own band training cut-off	Own band hold-out cut-off	Number of cases	Own band training cut-off	All band hold-out cut-off	
Accepted	2755	**76.19%**	75.97%	4069	**70.83%**	71.74%	**5.36%**
All case	4069	**72.16%**	72.13%	4069	**72.16%**	72.13%	

Table 5
Original data: Re-weighted logistic model with weights of evidence

Comparison 1: Area under ROC							
Predicting model	Training sample cases	Own band hold-out		All-applicant hold-out			Accept analysis delusion
		Number of cases	ROC area	Number of cases	ROC area		
Accepted	5413	2755	**0.7875**	4069	**0.7765**		0.0110
All case	8139	4069	**0.7837**	4069	**0.7837**		
Comparison 2: Percentage correctly classified							
Predicting model	Own band hold-out prediction			All-applicant hold-out prediction			Accept analysis delusion
	Number of cases	Own band training cut-off	Own band hold-out cut-off	Number of cases	Own band training cut-off	All band hold-out cut-off	
Accepted	2755	**76.15%**	76.04%	4069	**71.25%**	71.34%	**4.90%**
All case	4069	**72.16%**	72.13%	4069	**72.16%**	72.13%	

on an all application sample (Table 4 Comparison 2). (3) The re-weighted model gave a similar performance to the accept-only model when tested on the all-applicant sample (with the training sample cut-offs): 71.25% correct (Table 5 Comparison 2) compared with 70.83% (Table 4 Comparison 2), respectively. (4) Using an accept-only hold-out sample with accept-only cut-offs considerably over emphasizes the performance of the re-weighted model compared with a hold-out of all applications: 76.19% correct compared with 71.25% respectively (Table 5 Comparison 2).

7. How gains from reject inference depend on the rejection rate

In order to investigate the extent to which the efficacy of reject inference depends on the rejection rate, a new accept–reject model was estimated. This model then pro-

vided scores by which applicants' acceptance status could be varied. The new model required some of the data set to be dedicated to the accept–reject model and the rest to be dedicated to the good–bad models. The variables used to build each of these two types of model differed by an arbitrary selection such that each model had some variables that were not included in the other. We chose to estimate the accept–reject model with the 2540 Scottish cases in the data, and to estimate the good–bad model with the 9668 English and Welsh (hereafter English) cases in the data.

Typically, the accept–reject distinction would arise from a previous and perhaps somewhat obsolete credit-scoring model that distinguished good applicants from bad. It may also reflect some over-riding of credit scores by those using such a model. In setting up Scottish accept–reject and English good–bad models, the national difference in the data used for the two models appears as a metaphor for the inter-temporal difference that would separate the observations used to build two successive models. The exclusion of some Scottish variables in the development of the English model may be considered to represent, in part, the process of over-riding the acceptance criteria provided by the Scottish model. The exclusion of some English variables in the development of the Scottish model represents the natural tendency of new models to incorporate new variables not previously available. The progress of time also facilitates the incorporation of more variables by providing more cases and thereby permitting more variables to enter significantly.

A two-part procedure was used to select variables for the Scottish model. To retain the character of the data-supplier's original acceptance model an eligible pool of variables was identified by three stepwise (backward Wald) regressions using Scottish, English, and UK cases. In these regressions the data supplier's acceptance status was the dependent variable. An explanatory variable that survived in any one of the three equations was deemed to have possibly influenced the acceptance by the data supplier. The eligible variables were then used to model good–bad behavior in Scotland in a backward stepwise procedure that eliminated further variables.

The variable set for the good–bad model, to be parameterized on English data, was determined with a backward stepwise regression using English data, starting with all variables available to the English cases. A few scarcely significant variables common to the English and Scottish variable sets were then eliminated from one or the other to increase the distinctiveness of the two regressor lists. Table 6 indicates which of the original 46 variables were selected for each equation. Also indicated are the number of coarse classes assigned for each variable as well as the minimum frequencies. The former indicates the number of dummy variables required in total (the sum of classes minus one); the latter indicates possible sensitivity to individual cases.

The English data was scored using the variable set and estimated parameters derived from the Scottish model, and then collected into five bands according to this score. Table 7 shows the proportion of good cases in each of these non-cumulative bands and demonstrates a broad variety of performance, varying from just under 90% good in the first quintile to half that rate in the last. Each of these bands had training and hold-out cases determined by proportional stratified random sampling whereby in each band a third of good cases and a third of bad cases were randomly allocated to the hold-out sample. This sampling design was adopted to enhance

Table 6
Variables included in the accept–reject and good–bad equations

Variable description	Good–bad equation	Accept–reject equation	Classes	Minimum frequency
Time at present address		✓	8	281
B1		✓	4	242
Weeks since last county court judgement (CCJ)		✓	6	244
B2		✓	5	324
B3	✓	✓	6	453
Television area code	✓	✓	5	26
B4	✓	✓	6	496
Age of applicant (years)	✓	✓	6	201
Accommodation type	✓	✓	5	180
Number of children under 16	✓	✓	6	130
P1	✓	✓	3	377
Has telephone	✓	✓	3	1883
P2	✓	✓	6	611
B5	✓	✓	4	239
B6	✓	✓	5	320
P3	✓	✓	4	516
B7	✓		6	1108
B8	✓		6	407
B9	✓		6	1443
Type of bank/building society accounts	✓		6	188
Occupation code	✓		6	129
P4	✓		6	1108
Current electoral roll category	✓		5	458
Years on electoral roll at current address	✓		6	458
B10	✓		6	403
P5	✓		3	379
B11	✓		6	324
B12	✓		4	1163
B13	✓		4	1291
Number of searches in last 6 months	✓		4	406

Bn = bureau variable n; Pn = proprietary variable n; ✓ denotes variable is included.

comparability of corresponding hold-out and training cases and to retain the pattern of behavior in successive bands.

Finally, the bands were accumulated with each band including the cases of those bands previously above it. These are the bands used in subsequent analysis. Each band represents a possible placement of an acceptance threshold with the last representing a situation where all applicants are accepted.

For all of the bands the same coarse classification is used as in investigation one, but for each band separate weights of evidence were calculated for each variable for each of the five bands.

8. Banded results

In Table 9 the re-weighted results were calculated from an accept–reject model (to give the weights) followed by a good–bad model. The same variable sets were used in

Table 7
Sample accounting

Cases not cumulated into English acceptance threshold bands to show good rate variety

	All sample case			Good rate (%)	Training sample cases			Hold-out sample cases		
	Good	Bad	Total		Good	Bad	Total	Good	Bad	Total
Band 1	1725	209	1934	**89.2**	1150	139	1289	575	70	645
Band 2	1558	375	1933	**80.6**	1039	250	1289	519	125	644
Band 3	1267	667	1934	**65.5**	844	445	1289	423	222	645
Band 4	1021	912	1933	**52.8**	681	608	1289	340	304	644
Band 5	868	1066	1934	**44.9**	579	711	1290	289	355	644
English	**6439**	**3229**	**9668**	66.6	4293	2153	6446	2146	1076	3222
Scottish	1543	997	2540	60.7						
Total	**7982**	**4226**	**12208**	65.4						

Cases cumulated into English acceptance threshold bands for analysis

	English sample cases			Good rate (%)	Training sample cases			Hold-out sample cases		
	Good	Bad	Total		Good	Bad	Total	Good	Bad	Total
Band 1	1725	209	1934	**89.2**	1150	139	**1289**	575	70	**645**
Band 2	3283	584	3867	**84.9**	2189	389	**2578**	1094	195	**1289**
Band 3	4550	1251	5801	**78.4**	3033	834	**3867**	1517	417	**1934**
Band 4	5571	2163	7734	**72.0**	3714	1442	**5156**	1857	721	**2578**
Band 5	6439	3229	9968	**66.6**	4293	2153	**6446**	2146	1076	**3222**

both models for three reasons. First, this is representative of a typical application (see above). Second, we reduce the type of bias noted by Hand and Henley (1993/4). Third, if the variables of the original accept–reject model were used the model would be almost perfectly fitted yielding weights of 1 and infinity.

Comparison 1 in Tables 8 and 9 show our results using AUROC as a performance measure and Comparison 2 in these tables show our results using percentages correctly classified. These tables show weights of evidence results. Apart from showing that the scope for any improvements in performance increased as the cut-off in the original model is raised, as was shown in an earlier paper, these tables indicate many new findings. First, comparing Comparison 1 column 6 in Tables 8 and 9 (where the hold-out relates to a sample of all applicants) shows that the use of re-weighting reduces predicted performance compared with an unweighted model. Furthermore, the deterioration is greater for bands 1 and 2 than for bands 3 and 4. Generally, it seems the higher the cut-off score in the original accept–reject model the greater the deterioration caused by re-weighting.

Second, comparing the performance when tested on a hold-out sample from the accepted only (i.e. for each band separately) with that found when using a hold-out sample for all applicants (Comparison 1 in Tables 8 and 9, column 4 with column 6) shows that the former is overoptimistic in its indicated result. This is true for the unweighted model and for the model with re-weighting. For example, if the original accept–reject model had a high cut-off (band 1) and the analyst used these accepts to build and test a model, the indicated performance would be an AUROC of 0.8654 whereas the performance on a sample representative of all

Table 8
Band analysis: Simple logistic model with weights of evidence

Comparison 1: Area under ROC

Predicting model	Training sample cases	Own band hold-out		All-applicant hold-out		Accept analysis delusion
		Number of cases	ROC area	Number of cases	ROC area	
Band 1	1289	645	**0.8654**	3222	**0.7821**	**0.0833**
Band 2	2578	1289	**0.8249**	3222	**0.7932**	**0.0317**
Band 3	3867	1934	**0.8175**	3222	**0.8009**	**0.0166**
Band 4	5156	2578	**0.8108**	3222	**0.8039**	**0.0069**
Band 5	6446	3222	**0.8049**	3222	**0.8049**	

Comparison 2: Percentage correctly classified

Predicting model	Own band hold-out prediction			All-applicant hold-out prediction			Accept analysis delusion
	Number of cases	Own band training cut-off	Own band hold-out cut-off	Number of cases	Own band training cut-off	All band hold-out cut-off	
Band 1	645	**89.30%**	89.77%	3222	**70.20%**	72.56%	**19.10%**
Band 2	1289	**83.40%**	83.86%	3222	**70.58%**	72.75%	**12.82%**
Band 3	1934	**79.21%**	79.42%	3222	**71.97%**	73.49%	**7.24%**
Band 4	2578	**75.37%**	75.56%	3222	**72.47%**	73.81%	**2.90%**
Band 5	3222	**73.65%**	73.49%	3222	**73.65%**	73.49%	

Table 9
Band analysis: Re-weighted logistic model with weights of evidence

Comparison 1: Area under ROC

Predicting model	Training sample cases	Own band hold-out		All-applicant hold-out		Accept analysis delusion
		Number of cases	ROC area	Number of cases	ROC area	
Band 1	1289	645	**0.8483**	3222	**0.7374**	**0.1109**
Band 2	2578	1289	**0.7509**	3222	**0.7104**	**0.0405**
Band 3	3867	1934	**0.8034**	3222	**0.7920**	**0.0114**
Band 4	5156	2578	**0.8017**	3222	**0.8036**	**−0.0019**
Band 5	6446	3222	**0.8049**	3222	**0.8049**	

Comparison 2: Percentage correctly classified

Predicting model	Own band hold-out prediction			All-applicant hold-out prediction			Accept analysis delusion
	Number of cases	Own band training cut-off	Own band hold-out cut-off	Number of cases	Own band training cut-off	All band hold-out cut-off	
Band 1	645	**88.37%**	88.53%	3222	**69.77%**	68.96%	**18.60%**
Band 2	1289	**80.45%**	80.92%	3222	**68.56%**	67.60%	**11.89%**
Band 3	1934	**79.42%**	79.42%	3222	**72.38%**	72.94%	**7.04%**
Band 4	2578	**75.68%**	75.80%	3222	**72.84%**	73.74%	**2.84%**
Band 5	3222	**73.65%**	73.49%	3222	**73.65%**	73.49%	

applicants would be 0.7821 (Table 8 Comparison 1). The difference of 0.0833 is indicative of the error that would be made and we call this 'accept analysis delusion'. Val-

ues of this delusion are shown in column 7 in Tables 8 and 9. Notice that the size of the delusion is positively and monotonically related to the height of the cut-off in the original accept–reject model. Similar results are gained when dummy variables are used.

Our results using the percentage correctly classified are shown in Comparison 2 of Tables 8 and 9. Since an analyst would use the hold-out sample merely to test a model whose parameters (including the cut-off) were calculated from a training sample, one can see from columns 3 and 6 that the size of the delusion is substantial at cut-offs that equate expected and actual numbers of bads in the training band. For example, with a high cut-off (band 1) in the original accept–reject model the delusion is 19.10% of cases in both the unweighted and weighted models.

Third, column 6 of Comparison 2 in each of Tables 8 and 9 indicates the modest scope for improved classification by using information about the good–bad behavior of rejected applicants. Each result in that column indicates classification performance over applicants from all bands when parameters and cut-offs are taken from the particular band. In particular, the cut-off is taken such that predicted and actual numbers of goods in the training sample are equal. In this way the chosen cut-off reflects in part the band's own good–bad ratio, and takes no account of the all-applicant good–bad ratio. As we move from the low risk Band 1 to the higher risk bands below it we observe classification performances that approach that which is possible when no applicant is rejected. In Table 8, for example, the maximum scope for improved classification is only 3.45% (73.65%–70.20%). At best reject inference can but close this gap by producing better regression coefficients and/or by indicating better cut-off points.

Fourth, column 7 of Comparison 2 in each of Tables 8 and 9 suggests a negligible scope for reject inference to improve classification performance were the population good–bad rate to be actually known. In that column each band reports classification where each applicant is scored using regression coefficients arising from estimation in that band's training sample. However, the cut-off score is that which will equate the number of predicted bads among all applicants with the actual number of bads in the hold-out sample of all applicants. In this way each band's cut-off is determined by a good sample-based indication of the good–bad ratio for the whole population of applicants. As we move from the low risk Band 1 to the higher risk bands below it we see a maximum scope for improved classification of only 0.83% (73.49%–72.56%). Indeed for all but the top two bands there is no scope for improvement at all. The negative scope for improvement in Band 4 (73.49%–73.81%) must be seen as a reflection of sample error and indicates thereby how precariously small is even the improvement potential for band 1.

Of course, knowledge of the population good–bad ratio required to generate the results of column 7 in Comparison 2 is unlikely to be available, and so column 6 remains the likely indication to an analyst of the scope for reject inference to improve classification. However, since the scope for improvement all but vanishes in the presence of a suitable cut-off point, there seems correspondingly negligible potential benefit from the removal of bias or inefficiency in the estimation of regression coefficients used to score the applicants.

Finally, turning to the actual classification performance when re-weighting is used to attempt improvement in the estimation of regression coefficients, corresponding elements in column 6 of Comparison 2 of Tables 8 and 9 indicate very small improvements for Bands 3 and 4 and worse performances in Bands 1 and 2. For example, in Band 1 the performance of the re-weighted model is 69.77% compared with 70.20% for the unweighted model, yet in Band 4 the corresponding performances are 72.84% and 72.47%, respectively. An interesting comparison feature of the re-weighting procedure is shown by comparing Table 8 column 7 with Table 9 column 7. Table 9 column 7 presents a relatively large scope for improved performance even in the presence of a suitable cut-off that reflects knowledge of the population good–bad ratio. The potential for improvement is 4.43% (73.49%–68.96%). Therefore, while re-weighting undermines predictive performance by a minimal amount without such knowledge, it appears to undermine ability to deploy such information. Again these results were found when dummy variables were used instead of weights of evidence.

9. Extrapolation results

The foregoing discussion has demonstrated relatively little potential for improved regression coefficients but indicates considerable scope for using the population good–bad ratio to advantage. Extrapolation is mainly an attempt to obtain a good indication of that ratio. Rejects are first classified as good or bad by using a good–bad model parameterized using the training accept-only sample together with cut-offs that equalize the actual and predicted number of bads in the training sample of a particular band. These predictions are then combined with the actual good–bad values observed in the band, and an all-applicant model is calculated. This second model can hardly be expected to produce very different coefficients, so any scope for improvement will arise out of the application of a cut-off that reflects the good–bad ratio imputed for the all-applicant sample.

Table 10 shows that extrapolation gave a virtually identical predictive performance compared with a model estimated only for the accepts. This is roughly true for every band. With dummy variables the results are almost consistently better al-

Table 10
Band analysis: Extrapolation percentage correctly classified

All-applicant hold-out sample using training sample cut-off points					
Predicting model	Number of cases	Weights of evidence predictions		Dummy variable predictions	
		Simple logistic	Logistic with extrapolation	Simple logistic	Logistic with extrapolation
Band 1	3222	70.20%	**69.80%**	68.65%	**68.56%**
Band 2	3222	70.58%	**70.20%**	69.46%	**69.58%**
Band 3	3222	71.97%	**71.79%**	72.13%	**72.35%**
Band 4	3222	72.47%	**72.63%**	72.91%	**73.34%**
Band 5	3222	73.65%	**73.65%**	74.24%	**74.24%**

beit by a small amount. With weights of evidence the results seem very slightly worse when using extrapolation. However, that result might be reversed were the weights of evidence to be recalibrated using the imputed values of good–bad performance as in principle they should have been. The small margins of potential benefit indicated provide but a hint of what further research might indicate.

10. Conclusion

The analysis of reject inference techniques discussed above benefits from availability of a data set that permits the results of reject inference to be assessed in light of the actual repayment performance of "rejected" cases. The data set reflects a situation in which virtually no applicant was rejected in order for the data supplier to infer the character of the population of all applicants. The virtual absence of actual rejection in the supplied data has permitted consideration of both very high and low notional acceptance thresholds.

Unfortunately, neither an actual accept–reject score for each applicant nor the underlying model for determining it was available. Nevertheless availability of the accept–reject status that the data supplier would normally implement for each applicant has permitted an explicit and realistic accept–reject model to be fabricated. While this model does not reflect actual experience, it provides an explicit and plausible basis for inferring whether applicants might have been accepted.

One very clear result is the extent to which measures of predictive performance based on a hold-out sample of accepted applicants are liable to be misleadingly optimistic. This might have been expected in cases where the good–bad ratio is high, but the results presented here provide an empirical indication of the possible extent of error.

The other analytical findings seem quite plain. (1) Even where a very large proportion of applicants is rejected, the scope for improving on a model parameterized only on those accepted appears modest. Where the rejection rate is not so large, that scope appears to be very small indeed. That result is consistent with the data originally provided concerning the actual acceptance status of applicants and with the banded analysis that deploys a notional acceptance status. (2) Reject inference in the form of re-weighting applicants within a training sample of accepted cases and adopting a cut-off point based on those accepted cases appears to perform no better than unweighted estimation. In fact where the rejection rate is high, results appear to be quite noticeably worse. (3) Re-weighting appears to impede useful application of knowledge about the good–bad rate prevailing in the population, but without providing any compensating benefit. (4) Reject inference in the form of extrapolation appears to be both useless and harmless. It tends to leave regression coefficients unchanged, but the indication it provides about the population's good–bad rate seems to be inadequately accurate to provide benefit in spite of being informed by observed attributes of rejected applicants.

Useful implementation of reject inference seems to depend on accurate estimation of the potential good–bad ratio for the population of all applicants. Simple applica-

tion of that ratio then seems indicated. More elaborate tweaking of a vast set of coefficients does not seem to promise much potential benefit on the basis of the findings presented here.

References

Ash, D., Meester, S., 2002. Best practices in reject inferencing. In: Presentation at Credit Risk Modelling and Decisioning Conference. Wharton Financial Institutions Center, Philadelphia.

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

Boyes, W.J., Hoffman, D.L., Low, S.A., 1989. An econometric analysis of the bank credit scoring problem. Journal of Econometrics 40, 3–14.

Demster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data. Journal of the Royal Statistical Society B 39, 1–38.

Hand, D.J., Henley, W.E., 1993/4. Can reject inference ever work? IMA Journal of Mathematics Applied in Business and Industry 5, 45–55.

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

Heckman, J.J., 2001. Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. Journal of Political Economy 109, 673–748.

Hsai, D.C., 1978. Credit scoring and the equal credit opportunity act. The Hastings Law Journal 30, 371–448.

Joanes, D.N., 1993/4. Reject inference applied to logistic regression for credit scoring. IMA Journal of Mathematics Applied in Business and Industry 5, 35–43.

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

Meester, S., 2000. Reject Inference for Credit Scoring Model Development Using Extrapolation. Mimeo., CIT Group, New Jersey.

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

Thomas, L.C., Edelman, D.E., Crook, J.N., 2002. Credit scoring and its applications. In: Monographs on Mathematical Modelling and Computation. Society for Industrial and Applied Mathematics, Philadelphia.