LEAN MODELS AND REJECT INFERENCE

by

John Banasik and Jonathan Crook

University of Edinburgh

Corresponding author: John Banasik, Credit Research Centre, School of Management, University of Edinburgh, 50 George Square, Edinburgh EH8 9JY
John.Banasik@ed.ac.uk

Abstract

Credit scoring models are normally built using only applicants who have been previously accepted for credit. Such non-random sample selection may produce bias in estimated model parameters and accordingly model predictions of repayment performance may not be optimal. Previous empirical research suggests that omission of rejected applicants has a detrimental impact on model estimation and prediction., This paper explores the extent to which the number of included variables influences the efficacy of a commonly used reject inference technique, reweighting. Analysis benefits from availability of a rare sample where virtually no applicant was denied credit. The general indication is that the efficacy of reject inference is little influenced by either model leanness or interaction between model leanness and the rejection rate that determined the sample. However, there remains some hint that very lean models may benefit from reject inference where modelling is conducted on data characterized by a very high rate of applicant rejection.

Introduction

One of the main limitations of most credit scoring models is that they are used to assess the creditworthiness of all applicants, yet they are formulated and calibrated on the basis only of applicants previously judged good enough to be granted credit. The effectiveness of such models depends on good applicants differing from bad applicants with respect to their attributes such as age or income that are deployed by such models to predict repayment behaviour. Such effectiveness also depends on good applicants not differing among themselves with respect to how well their repayment behaviour is predicted by the structure and parameters of the model. Otherwise the disproportionate omission of bad borrowers among the rejected applicants will result in biased predictions with respect to repayment behaviour. Reject inference techniques attempt to obviate such possible bias that arises from calibrating models in the absence of rejected applicants, and involve resort to characteristics about those applicants. Unfortunately, the key missing characteristic of such applicants is the repayment behaviour that would have emerged had they been accepted. Any corrective influence of reject inference techniques will, therefore, be at best partial.

Hand and Henley[1] have argued that reject inference is unlikely in principle to yield improved model effectiveness. Essentially, if there is a single model that governs good borrowers and bad equally, then its parameters may be reliably inferred on the basis of the better borrowers that have been accepted. Absence of rejected applicants may result in inefficient discernment of model parameters but not biased estimates of them. If a single model does not govern accepted applicants and rejected applicants equally well, then it is by definition misspecified and in principle should be scrapped

in favour of one that does govern all. Reject inference thus seems an ill-considered attempt to remedy the defect of models that are fatally flawed by misspecification bias. A good model should therefore have no scope for benefiting from reject inference.

The practical justification for reject inference is that, in the absence of observed repayment behaviour for rejected applicants, the extent to which there exists a single model that governs all is difficult to discern. Accordingly, a model built upon accepted applicants only is presumed universally applicable, and reject inference is applied as a failsafe against the possibility that it is not. At worst it is useless and at best it reduces bias but does not remove it. An empirical assessment of reject inference will indicate two features of credit scoring models, the difficulty in achieving a universally applicable model and the corrective influence of an adopted reject inference technique. Inevitably any conclusions drawn will be the tentative result of a given data set and model.

Relatively little has been published that empirically assesses particular reject inference techniques. Meester[2] considered two methods of extrapolation and found some modest scope for prediction improvement. Banasik et. al.[3] considered bivariate probit for sample selection, using a sample with virtually no rejected applicants, and concluded that this approach enabled only minimal improvement compared with prediction on the basis of a model callibrated on accepted applicants only. Crook and Banasik[4] evaluated another two techniques, augmentation (often called re-weighting) and extrapolation, with the same data set and a virtually identical framework of reference and found modest but distinct scope for the corrective influence of reject inference techniques. However, extrapolation was found to be useless but harmless

while augmentation provided a distinctly inferior predictive performance compared with a sample of applicants that might normally be accepted.

This study revisits the findings of the Crook and Banasik[4] (forthcoming) study to consider the extent to which the breadth of available variables may have undermined the scope for reject inference to improve predictive performance. The model used to predict repayment performance deployed 26 variables each with three to six coarse categories – most typically six. This volume of detail might conceivably render prediction more robust that it would otherwise be. Consider repayment behaviour explained by only one variable comprising six coarse categories represented by five binary variables and a constant term, and consider parameters estimated from a sample that excludes 80% of applicants previously deemed not to be creditworthy. Assume that the model reflects rejected applicants in terms of its structure but not its estimated parameters. This model might nevertheless predict repayment behaviour among rejected applicants significantly better than random selection. A model with many such variables might to some extent be regarded as a concatenation of simpler models with a tendency for errors to average themselves out when applied to rejected applicants. Such a model might tend to benefit less from reject inference than a lean one.

The term "lean" model is here used to denote a model with "few" explanatory variables. Given the limited sample size, between 1289 and 6446 cases, and potentially large number of binary variables used, up to 105, "lean" models also tend to be models characterized by few insignificant coefficients. This will imply an unfortunate tendency for prediction loss from having only few variables being offset by efficiency gains from the removal of insignificant variables. Such distortion is the

cost of data scarcity, but might be presumed mainly to effect the reliability of parameter estimates rather than the predictions generated.

The next section of this paper outlines briefly the character of the augmentation reject inference technique that will be considered here. The following section indicates the process by which an index for acceptability was generated so that the efficacy of reject inference could be assessed for different acceptance rates. The next section presents results for a model that deploys 26 variables comprising a total of 105 coarse categories represented by binary variables. In this section the efficacy of augmentation is examined in a very similar way to that appearing in Crook and Banasik[4] except that binary variables are used instead of weights of evidence variables. For both of these types of reject inference a comprehensive range of models with varying leanness are then evaluated. The concluding section summarizes and discusses these results.

Augmentation

The main rationale for augmentation probably depends on a presumption that no single parameter set governs equally the better applicants that tend to be accepted and the worse applicants that tend to be rejected. These techniques thus tend to seek a set of "average" parameters that will focus attention away from better applicants toward more typical applicants. In this way parameters may tend to reflect the character of more marginal applicants whose repayment behaviour is most uncertain.

Augmentation, probably the most widely used reject inference technique, involves weighting accepted applicants in such a way as to synthesize a sample in which rejected applicants are fully represented. The following method adopted here is but

one of several similar variants adopted by practitioners. First, an Accept-Reject (AR) model predicts whether an applicant has been accepted or rejected on the basis of a particular set of explanatory variables, and on this basis all applicants, accepted or not, are assigned a score.

In order to appreciate the use to which these scores are put, now consider for the moment that there exist ranges of equivalent scores, so that cases can be assigned to intervals as illustrated in Table 1. In this table each of these ranges of equivalent scores is represented by a interval wherein there are both accepted cases and rejected cases. Among the accepted cases it is possible to distinguish between those whose repayment behaviour was bad (hereafter "bads") and those for whom it was not (hereafter "goods"). Specifically, accounts transferred for debt recovery within 12 months of the credit being first taken were defined as bad, all other accounts were defined as good. In each interval the acceptance rate can be discerned and this may be taken as an applicant's probability of being accepted within a given score interval. Assume that within each interval accepted and rejected applicants are equally likely to have good repayment performance. Then weighting the accepted cases in each interval by the inverse of the that interval's acceptance probability will give each interval the relative weighting that would prevail were all rejected applicants to have been accepted.

Table 1: Re-weighting

Interval (j)	Number of of Goods	Number of of Bads	Number of Accepts	Number of Rejects	Interval Weight
1	n_{g1}	n_{b1}	$n_{A1} = n_{g1} + n_{b1}$	n_{R1}	$(n_{R1} + n_{A1}) / n_{A1}$
2	n_{g2}	n_{b2}	$n_{A2} = n_{g2} + n_{b2}$	n_{R2}	$(n_{R2} + n_{A2}) / n_{A2}$
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

-	-	-	-	-	-
n	n_{gn}	n_{bn}	$n_{An} = n_{gn} + n_{bn}$	n_{Rn}	$(n_{Rn} + n_{An}) / n_{An}$

The arbitrary distortion implied by the creation of intervals of equivalent scores can be eliminated by increasing the number of intervals such that there is one for every single observed score. Where the scores are the probabilities arising from a logistic regression model, for example, the process can be applied more simply and comprehensively by attributing to each accepted case a weight that is the inverse of its probability of acceptance according the AR model. These weights can then be used to weight cases in a GB model that predicts among accepted cases whether the case will be good or bad. This second model should have some of the character of the GB model that would have been feasible to estimate were the repayment of rejected applicants to have been known.

Two remarks of this procedure deserve mention. First, bias will result should the set of explanatory variables used for the initial AR model include explanatory variables not used in the subsequent GB model. For criticism along these lines, see Hand and Henley[3,5] who build on the work of Little and Rubin[6]. An account of this criticism also appears in Crook and Banasik[4] (forthcoming). In the analysis to follow the AR and GB models share a common set of explanatory variables even though acceptance designation is known to have depended on variables outside those used in the GB model. In this respect the model realistically incorporates imperfect knowledge of previous models likely to be present in many, if not most, applications, and it avoids the bias alluded to above at the expense of accepting some model misspecification bias instead. Secondly, this procedure is not feasible where the previous GB model used to reject applicants is perfectly known, mechanistically applied without

overrides, and is adopted as the AR model used to determine weights. Even were the logistic regression equation feasible to compute, it would generate unit probabilities for all accepted cases and hence undefined weights.

Acceptance Bands

The present analysis benefits from a rare sample of applicants from which virtually no one was rejected. The credit supplier occasionally absorbs the cost of accepting applicants that would normally be rejected so as to have a data base for credit scoring models that can safely avoid any resort to reject inference techniques. Table 3 makes clear that applicants with very poor credit histories were accepted and inspection of individual cases confirms the presence of applicants with very poor credit histories. Accordingly, the rejection of a negligibly small proportion of applicants should have no influence on the conclusions presented here. The credit supplier also indicated which applicants in the data set would normally have been rejected.

Preliminary analysis indicated that there was extraordinarily little scope for reject inference. Models built on those applicants that would normally be accepted predicted repayment for all applicants as well as those calibrated on all applicants. The proximity to zero scope was startling! The presumed cause of this result was that even normally very risky applicants would be accepted. The two-thirds of applicants normally accepted included nearly 30% who were "bad". Models built upon such accepted applicants would already incorporate insights into the characteristics of very bad applicants, and such insight would be denied models built by more cautious credit

suppliers. The extent to which the scope for reject inference depended on the acceptance rate thus became a major concern.

Little useful was known about how acceptance of applicants was normally determined except that most of the relevant variables had been provided, so that varying the acceptance threshold would require fabrication of an AR model. Normally, the evolution of a firm's GB model reflects new insights and changing characteristics among applicants. Over the course of time as a GB model becomes obsolete it also becomes the AR process for determining the data set upon which a subsequent GB model is formulated and estimated. For this study nationality was used as a metaphor for time. The 2540 Scottish applicants were held out as the sample upon which an AR model would be formulated and estimated. The remaining 9668 English and Welsh (hereafter English) applicants would then be the basis upon which GB models would be formulated and estimated as a basis for assessing the efficacy of reject inference. Further details are given in Crook and Banasik (forthcoming).

The variables that were included in each equation after the selection process are shown in Table 2. Among other things, the process of denying AR variables to the GB model reflects the possible over-riding of a credit scoring model, implying notional accommodation of additional variables Coarse classification was not supplied by the data supplier. They were based on a training sample used for the preliminary analysis in which the credit supplier's accept-reject designation was used. These seemed adequately robust to be used in all subsequent analysis.

Table 2: Variables Included in the Accept-Reject and Good-Bad Equations

Variable Description	Good-Bad Equation	Accept-Reject Equation	Coarse Categories	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

English applicants were grouped into five bands of nearly equal size according to their AR score arising from the Scottish model. From each band two-thirds of applicants were assigned to training samples for estimating scoring models and cut-off points. Stratified random sampling determined that these samples would have virtually the same good-bad rate as did the holdout samples used to assess predictive performance.

Table 3 illustrates the remarkably broad range of repayment performance represented by these bands with nearly 90% classed as good in the top band and half that rate in the bottom band. For purposes of all further analysis, however, the bands were accumulated such that each included all the cases in the band above it. Each band then corresponded to a distinct possible sample of accepted applicants. The all-inclusive fifth band was the all-applicant sample that would be the benchmark against models based on less inclusive samples of accepted applicants could be assessed.

Table 3: Sample Accounting

Cases Not Cumulated into English Acceptance Threshold Bands to Show Good Rate Variety:

	All Sample Case			**Good**	Training Sample Cases			Hold-out Sample Cases		
	Good	Bad	Total	**Rate**	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**	Training Sample Cases			Hold-out Sample Cases		
	Good	Bad	Total	**Rate**	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	9668	**66.6%**	4293	2153	**6446**	2146	1076	**3222**

Full Model Analysis of Reject Inference

Analysis of the full English model comprising all 26 variables provides the framework for all subsequent analysis. For each band, variable coefficients were estimated using the band's training sample and a cut-off point was found that equalized the predicted

number of training sample bads with that actually observed in the training sample. The hold-out cases for each band were then scored and classified on the basis of these training sample results. For example, Table 4 indicates that 87.91% of the 645 Band 1 hold-out cases were correctly classified in this way. Within the hold-out samples the predicted number of bads will generally not equal the actual number precisely because of sampling error. The cut-off point could be revised to remedy this, but this would tend to exaggerate the predictive potential of the model, because it would permit outcomes to condition predictions thereof. Nevertheless, Table 4 also reports results implied by standardizing hold-out sample performance in this way, mainly to confirm that the extent of exaggeration implied by this expedient is modest. For example, in Band 1 the increase from 87.91% to 88.53% implies only 4 more of the 645 cases correctly classified as a result of this expedient, and in most other bands the difference is much less.

The more interesting results are the extents to which models generated from within each band can correctly classify the 3222 cases of the all-inclusive Band 5's hold-out sample. The extent to which they fall short of the classification achieved by the Band 5 model defines the scope for reject inference to improve classification. For example, the variable coefficients and cut-off score generated from Band 1's training sample correctly classifies 69.09% of these cases. That represents a scenario where a model has been developed from a data base comprising the 20% of all applicants previously thought to be best. The fully representative Band 5 produces a model that correctly classifies 74.39% of cases. Accordingly, the scope for reject inference to improve predictive performance would be 5.30% (i.e. 74.39 – 69.09).

The last column of Table 4 indicates the potential of models generated within each band to classify Band 5 holdout cases were only the Band 5 cut-off known. Coefficients are taken from each Band's training sample, but the cut-off is selected to equalize the predicted number of bads with the actual number in the Band 5 hold-out sample. That is equivalent to indicating the potential performance of a model based only on accepted applicants where the proportion of bads among all applicants is known. The extent to which a Band's model still falls short of the Band 5 standard is the extent to which coefficients generate an inferior ranking of applicants. For example, the scope for reject inference to assist the Band 2 model can be decomposed into 3.11% (i.e. 73.00 – 69.89%) from knowledge of the all applicant, all band cut-off and 1.43% (i.e. 74.43 – 73.00) from better ranking of applicants.

Table 4: Augmentation Relative Performance for Full 26-Variable Model

Comparison 1: Percentage Correctly Classified without Augmentation

	Own Band Hold-out Prediction			All-applicant Hold-out Prediction		
Predicting Model	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	87.91%	88.53%	3222	**69.09%**	**69.96%**
Band 2	1289	83.63%	84.48%	3222	**69.89%**	**73.00%**
Band 3	1934	80.30%	80.56%	3222	**72.25%**	**74.43%**
Band 4	2578	76.30%	76.57%	3222	**73.15%**	**74.92%**
Band 5	3222	74.39%	74.43%	3222	**74.39%**	**74.43%**

Comparison 2: Percentage Correctly Classified with Augmentation

	Own Band Hold-out Prediction			All-applicant Hold-out Prediction		
Predicting Model	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	87.91%	89.15%	3222	**68.53%**	**68.53%**
Band 2	1289	82.39%	83.86%	3222	**69.40%**	**72.44%**
Band 3	1934	79.52%	80.25%	3222	**71.51%**	**74.18%**
Band 4	2578	77.35%	77.42%	3222	**74.02%**	**74.61%**
Band 5	3222	74.39%	74.43%	3222	**74.39%**	**74.43%**

The Table 4 Comparison 2 provides similar results where coefficients have been derived using the re-weighting technique described above. Results are generally inferior to those from unweighted regression with the exception of Band 4 where classification seems to enjoy some spurious superiority to that achieved in Band 5 for the last column. Note that samples of 3222 governed by a probability of 70% are subject to a standard error of about .8% so that one-tail significance of .01 requires difference of roughly 2%. Results reported in this paper are intended primarily to indicate orders of magnitude and patterns thereof. In general the inferior tendencies for the first three bands in the penultimate column of Comparison 2 indicate that Augmentation achieves none of the available scope for improved classification where a high proportion of applicants have been rejected. The last column of Comparison 2 indicates that augmentation actually tends slightly to undermine the potential for making good use of knowledge about the proportion of bads in the population.

Lean Model Analysis of Reject Inference

The principles of analysis in the last section will be carried forward in considering the application of reject inference for models with different numbers of variables. For each Band and each number of variables the total scope for reject inference will be indicated and divided between potential for improved ranking of applicants and from better awareness of the proportion of bad applicants in the overall population. For each of the two types of reject inference the extent of achievement is considered both with respect to ranking and cut-off points. The analysis of the above section was replicated for 23 model sizes, one for each number of variables between 4 and 26.

The smallest models included the most significant variables. Beyond this criterion little arbitrary choice was required. Results of a forward stepwise logistic regression procedure agreed remarkably well with those of a backward procedure, so that determining which variable to add to successively larger models was otherwise virtually unambiguous. Given the number of coarse categories for each variables the smallest reasonable number of variables seemed to be 4. The possible number of distinct scores implied by a model is the product of the number of coarse categories for the variables it includes. Unless there are sufficient distinct scores classification performance will depend to a large extent on the arbitrary allocation of tied applicants around cut-off points. Influence of ties did not disappear altogether until there was 17 variables in the model, but with as few as 4 variables the amount of distortion on this account was acceptably small. In order to achieve the training sample cut-off in a situation of ties, the cut-off was located at the nearest location of adjacent differing scores. In order to achieve appropriate hold-out sample cut-offs a very small random number (enough to leave the first eight digits after the decimal unaffected) was added to each score.

Figure 1 illustrates the scope for reject inference models with each number of variables. Given the scope for sampling error to throw up spurious results the focus of attention should be on the smoothed results[(a)]. Both raw and smoothed results are provided in Figure 1 to assure the reader that this process does not conceal significant patterns. In other figures such reassurance would be distracting.

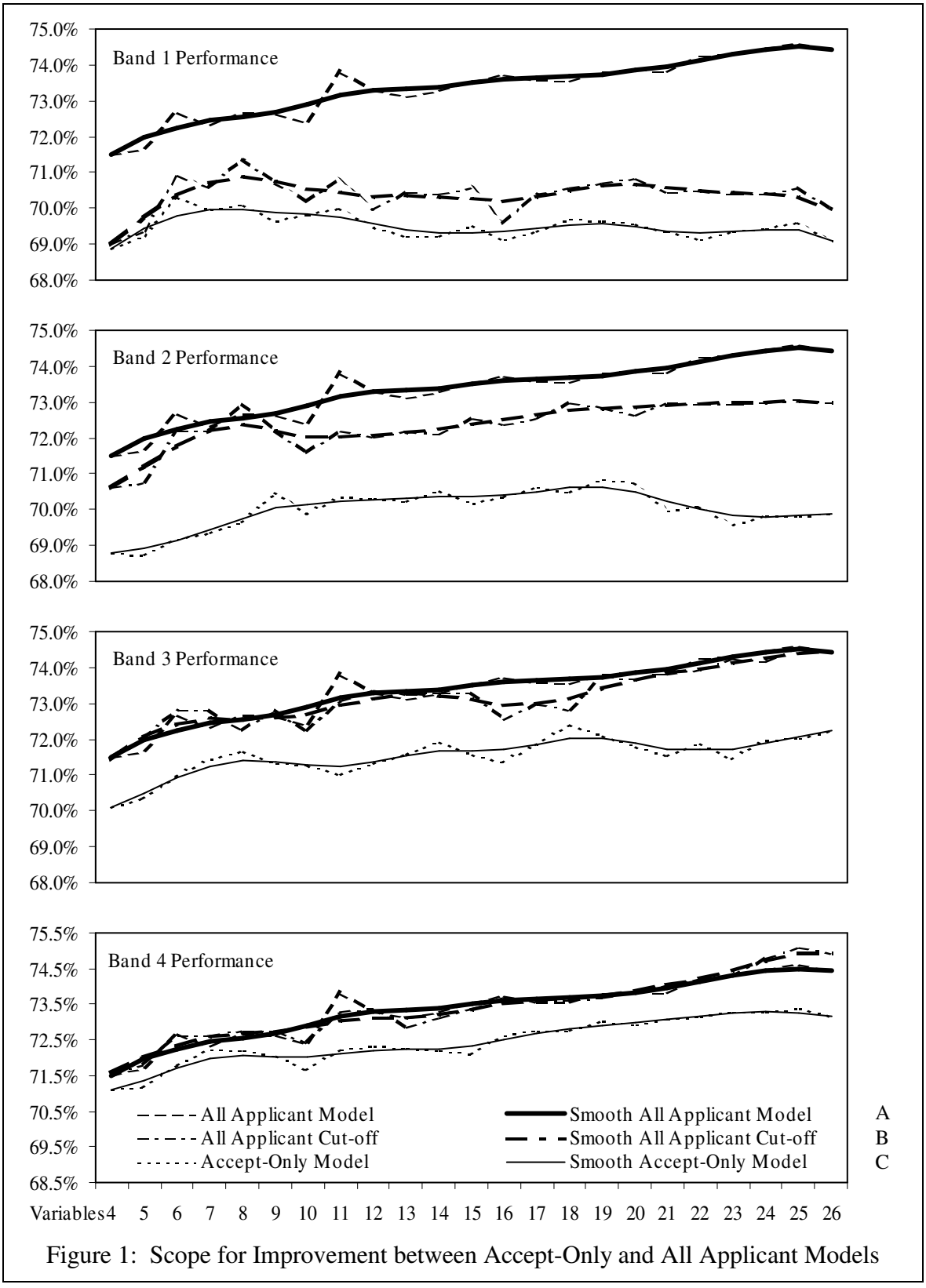

Figure 1: Scope for Improvement between Accept-Only and All Applicant Models

With 26 variables included, line A, the performance of the all applicant model using the all applicant, all band cut-off, has the same value as in Table 4, Comparison 1,

column 7, band 5: 74.43%. Line B, for band 1, the performance of the accept only model using the all applicant, all band cut-off, has the same value as Table 4, Comparison 1, column 7, band 1: 69.96%. Line C for band 1, the performance of the accept only model using the accept only cut-off has the same value as Table 4, Comparison 1, column 6, band 1: 69.09%. Corresponding interpretations hold for lines B and C for bands 2, 3 and 4. Therefore, as explained with regard to Table 4, the difference between lines A and C for, say, band 2, shows the total scope for reject inference to assist the band 2 model.

First, notice that the inclusion of decreasingly significant variables as model size increased accounts for the modest upward slope of all of the smoothed curves as well as their slight concavity. The absence of upward slope in the higher bands reflects the burden of adding insignificant variables to the explanation small samples. This will also account for the tendency of the slope to evaporate at the end of curves for all bands. The stepwise regression that determined the 26 variables suitable for English cases was estimated with all Band 5 cases, training and hold-out, and with a view to a 10% level of significance. Variable selection is an a priori condition for all Models and all bands, since this study does not include consideration of the effect of rejected applications on model formulation. Accordingly, when the full 26-variable model is estimated on training cases only, four of variables lose significance altogether. The 23 variable model is the largest fully significant model for Band 5.

The motivation for this study was an expectation that the scope for reject inference would decrease as the number of explanatory variables increased. That pattern is not evident. The scope for reject inference to improve performance can be partitioned into two parts. First the difference due to the ability to parameterise a model

representative of all applicants (band 5) and the corresponding cut-offs for all applicants (line A) and, on the other hand, the ability to parameterise a model using merely a sample from a particular band, but applying this to equate predicted and actual number of bads in the all applicant sample (line B). Thus (A-B) shows the effect of an improved ranking of applicants. The second portion is the difference between line B and, on the other hand, prediction from estimated parameters for a sample of accepts from the band with training sample cut-offs (line C). Thus (B-C) shows the effect, for any given band, of the ability to use the estimated model for that band with the all applicant sample, rather than the training sample only; that is the ability to detect an improved cut-off.

Examination of Figure 1 shows that the scope for reject inference, (A-C), if anything, appears to increase as more variables are included in the model. This may be the result of variable insignificance that would not prevail were the sample sizes larger, but there seems insufficient evidence to conclude that these lines would otherwise be converging.

The main indication from Figure 1 is that model leanness has little influence on the scope for rejection inference and on the nature of this scope. Table 3 demonstrated that in Band 1 the scope was relatively large and mainly to do with better ranking of applicants and little to do with awareness of an appropriate cut-off point. From Band 2 this situation was reversed; the scope diminished remarkably and what remained had mainly to do with awareness of an appropriate cut-off. These conclusions appear not to depend on the number of explanatory variables deployed.

One feature not evident from Figure 1 is that the similarity of performance between accept-only and all-applicant models does not reflect as close agreement as to which particular applicants are bad. The last column of Table 4 Comparison 1 implies a difference of 144 misclassifications between Band 1 and Band 5 and half of this difference relates to bads, since for this column actual bads must equal predicted bads. Table 5 indicates that of the 1076 predicted to be bad by each full model of 26 variables, only 758 are common to both bands 1 and 5. Accordingly, each model predicts 318 cases to be bad that are not predicted by the other, but the difference in correct predictions of bads is only 72. The Band 5 model is not simply finding more bads; it is finding many different bads. When only 4 variables are used only 797 bad cases are common to models using bands 1 and 5. With fewer variables there is greater commonality in the identities of the correctly predicted bads.

Table 5: Numbers of Bads Agreed upon by Different Band Models

	Models Estimated with 4 Variables					Models Estimated with 26 Variables				
	Band 1 Model	Band 2 Model	Band 3 Model	Band 4 Model	Band 5 Model	Band 1 Model	Band 2 Model	Band 3 Model	Band 4 Model	Band 5 Model
Band 1 Model	1076	895	786	796	797	1076	817	760	755	758
Band 2 Model		1076	921	898	899		1076	903	889	879
Band 3 Model			1076	1016	1016			1076	972	954
Band 4 Model				1076	1068				1076	1010
Band 5 Model					1076					1076

Having established that the scope for reject inference does not depend noticeably on model leanness, it remains to consider whether the effectiveness of reject inference depends on model leanness. Figure 2 confirms the findings of Table 4. In Figure 2, lines A,B and C have the same interpretation as in Figure 1. Line D shows performance using augmentation but the cut-off for the sample of accepts in a band.

Line E shows the performance using augmentation with the cut-off for the all applicant sample. The figure shows that augmentation undermines both performance and ability to exploit knowledge of the Band 5 cut-off. The four-variable model provides an interesting exception to this finding in that this very lean model benefits from augmentation in both respects. In many cases, however, the leaner models are worse in both respects than the full model.

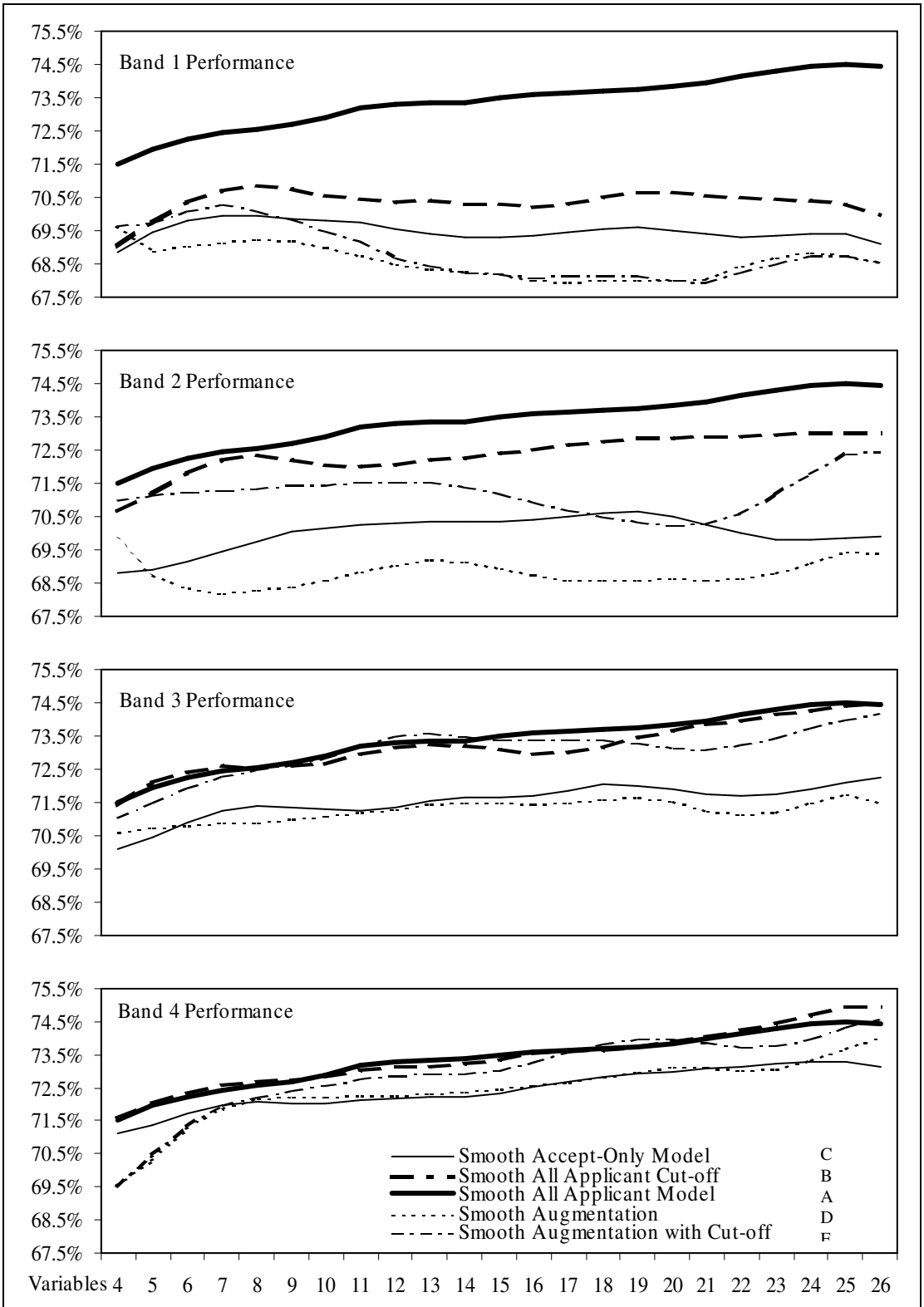

Figure 2: Model Performance for Models with and without Augmentation

Bands 1 and 2 provide the most noteworthy features. In Band 1 augmentation produces distinctly worse results than prediction in its absence once about eight variables are present in the model (line D is lower than line C) and from about a dozen variables knowledge of the Band 5 cut-off is of no use at all (lines D and E are virtually identical). In Band 2 augmentation without the Band 5 cut-off (line D) performs comparably poorly, undermining performance perhaps even more. However, the curve corresponding to augmentation using the Band 5 cut-off (line E) is drifting upward, so that augmentation "with Cut-off" outperforms no augmentation without it. However, that is of little practical use; either way, with or without Band 5 cut-off, one is better off without augmentation. The other bands represent situations where rejected applicants comprise 20% to 40% of all original applicants. In these Bands the scope for reject inference is small as is the damage augmentation does. While the damage is small and quite consistent, it does not approach anything resembling significant magnitudes.

Normally re-weighting does not introduce bias but may lose efficiency. In effect re-weighting replaces the influence of extremely good applicants with more influence of middling applicants. While that may arguably provide a more relevant focus of attention, it also tends to narrow the sample's range of observations, and the influence on estimation may be comparable to operating with less observations. Coefficients would not be biased as a result, but prediction error would tend to increase. To some extent Bands 1 and 2 are already deprived of observations, so that it is unfortunately difficult to distinguish interaction between high rejection rates and re-weighting from interaction between small samples and re-weighting.

Conclusion

The above analysis confirms the existence of scope for reject inference to improve predictions. Benchmark performance in the sample considered is not 0% correctly classified. Simply classifying every applicant to be good will imply 66% correctly classified. For the 34% left to account for the scope of 5.30% indicated above is not negligible progress. Clearly, the data set used here represents a situation in which rejecting applicants will lose information useful for subsequent modelling. What remains unclear is how much of the lost information about the repayment potential of rejected applicants can be recouped by deploying knowledge of their other attributes.

The simple reject inference technique considered here, augmentation, has generally done little harm and the extent to which harm is evident can to some extent be attributed to the aggravation of small-sample estimation that it implies. Nevertheless, the smallest training sample contains 1289 observations, so that the patterns that emerge do seem worthy of note, and the clearest pattern is that the technique does not achieve much good.

The focus of this paper is on the possible influence of model leanness. Models with few variables can often predict nearly as well as much more elaborate models, and this has been evident in the data set used here. Lean models will also tend to comprise highly significant variables, so that the net impact of particular attributes may reliably be discerned. The question addressed here is to what extent such models become sensitive to the characteristics of a particular sample, so that sample selection bias from rejected applicants might warrant and benefit from corrective attention. The general indication of the evidence presented is that neither the scope nor effectiveness of reject inference is much affected by this feature. On the contrary, there seems an

indication that scope for reject inference is positively related to model flabbiness. However, it is difficult to distinguish the influence of more variables from loss of significance for those variables as the source for this relationship, so it may simply be a sample-size effect.

The other major pattern is that interesting things happen in the higher bands and not in the lower bands. The small scope for improved performance in Bands 3 and 4 from reject inference and the virtual absence of scope for improved ranking of applicants in these bands suggests generally that reject inference is rewarding mainly when *most* applicants are rejected. In the higher bands there is clear scope for reject inference but no definite indication that augmentation has anything to offer empirically. Some possibilities are suggested, but it is difficult to establish to what extent this actually reflects the high rejection rates implied by these bands as opposed to the influence of small-sample anomalies and resort to insignificant variables. The scope for model leanness to interact with a high rejection rate to give prospects for reject inference are therefore at best murky. The overall impression is that such prospects exist but are slight. However, it is difficult to imagine how such prospects will be clarified empirically, since the data set used here was a rare instance of a repayment details of applicants with a very broad range of creditworthiness.

References

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

2 Meester, S, (2000). *Reject Inference for Credit Scoring Model Development Using Extrapolation.* Mimeo, CIT Group: New Jersey.

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

4 Crook, J N and Banasik, J L, (forthcoming). Does reject inference really improve the performance of application scoring models? *Journal of Banking and Finance*.

5 Hand DJ and Henley WE (1994) Inference about rejected cases in discriminant analysis. In: Diday E, Lechvallier Y, Schader M. Bertrand P and Buntschy B (eds) *New Approaches in Classification and Data Analysis*. Springer-Verlag: Berlin, pp 292-299.

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

Footnotes

[a] The end points of the smoothed results are the actual values, for the second and penultimate model they are centred three-point moving averages and for all other models they are three-point moving averages of three-point moving averages.