Credit Risk Scorecard Design, Validation and User Acceptance

- A Lesson for Modellers and Risk Managers

Edward Huang and Christopher Scott
Retail Decision Science, HBOS Bank, UK

1. Introduction

Credit risk scoring has gone a long way since Fair Isaac introduced the first commercial scorecard to assist banks in making their credit lending decisions over 50 years ago. It now becomes the cornerstone in modern credit risk management thanks to the advancement in computing technologies and availability of affordable computing power. Credit scoring is no longer only applied in assessing lending decisions, but also on-going credit risk management and collection strategies. Better designed, optimally developed and hence more powerful credit risk scorecard is a key for banks and retail finance companies alike to achieve competitive advantage in today's competitive financial services market under the tough economic environment with severe consumer indebtedness. Several books have been published which serve as a good introduction to credit management and scoring. [1-4]

Scorecard development methodology has evolved over the years too. Starting from the divergence based scorecard method used by FICO which dominated the industry for the first a couple of decades, it has now diversified into a spectrum of methodologies from which researchers, practitioners and consultancy firms may choose for their scorecard development, such as logistic regression, decision trees, mathematical programming, neutral network [5], genetic algorithm [6], survival analysis modelling [7-8], support vector machine [9] and graphical models [10-11], etc. Some seminal work has been done in applying double hurdle modelling technique to credit risk scoring problem [12]. Among all these, logistic regression is now the most commonly used method for scorecard building.

Compared to the on-going exploration of new modelling techniques for credit scoring, there has been much less focus (at least much less reported case studies) on the practical side of scorecard design, validation and user acceptance which is equally important to credit industry if not more. This paper illustrates and discusses a fairly common situation where newly developed credit risk scorecard may appear to perform better on the development sample and validation sample, but deteriorates significantly when being assessed on out-of-time sample or in real implementation.

We shall describe the problem in section 2, followed by more detailed investigation into the underlining root causes in section 3, and finally summarise the major findings and the general implications to both modellers and risk managers.

Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com

2. The Problem

Credit risk management can be broadly divided into front-end (or acquisition) risk management and back-end (or existing customer) risk management. The most important decision support tool for front-end approval/reject decision making is Application Scorecard (A-score) while that for the back-end customer risk management decision making is Behaviour Scorecard (B-score), both of them measure the likelihood of a customer becoming default over a certain period of time (i.e. performance window). As a common practice, performance window for A-score is chosen to be 12 – 24 months and that for B-score 6 – 18 months. In recent years, more and more banks are becoming more sophisticated and they started to build a suit of A-scores and B-scores to predict not only default propensity, but credit abuse, insolvency and over-indebtedness, etc. Because of the dynamic nature of retail finance products as a result of fierce competition and the population attitude and behaviour shifts, it became a norm that such scorecards are regularly reviewed and redeveloped. Many banks impose a compulsory redevelopment cycle such as 24 months.

Like any modelling process, a scorecard model should not only fit the development sample well (although not over-fitted) to gain confidence, but to demonstrate its robustness and stability through independent validation sample and out-of-time validation sample (or call it temporal validation). It is a common practice in the credit industry that scorecard modelling sample is split into model development sample and validation sample, usually 70% vs. 30%, while out-of-time validation sample is collected using another time reference point. Modellers and educated risk managers would only accept the new scorecard for implementation when both validations are satisfactory, and when there is strong evidence that the newly built scorecard will result in tangible business benefit.

Development, validation and on-going monitoring of A-score are somewhat more complex for the reason that the scorecard is intended to score all applicants (i.e. through-the-door population), but the performance data are only available to those who were approved. Although Reject Inference technique could be used in model development process to infer the Good/Bad outcome of an applicant if the person's application was actually rejected, scorecard performance monitoring are done based on accepted population only. Rightly or wrongly, many risk managers still rely on the evidence they see from Accept Only population to decide whether they are willing to accept the scorecard for implementation, given the efforts risk management team needs to put in to revise risk strategies in accordance to the new scorecard, particularly where life-time expected loss and projected life-time profitability are applied in making cut-off decisions. This view reflects the lack of conclusive evidence on the effectiveness of reject inference although it has become a widely accepted practice. Hand and Henley showed that the methods typically used in the industry are problematic and hence concluded that reliable rejection inference is impossible [13]. Chen assessed the effectiveness of Heckman's two-stage model with partial observability for

Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com

reject inference using complete information on both rejected and accepted loan applicants [14]. Chen concluded from the study that using this reject inference technique did not justify its marginal gain to credit classification although it is theoretically a sound reject inference method.

While the academic community has failed so far in establishing a reliable method for reject inference to compensate the sample bias in application scorecard development, the only convincing approach would be to allow some customers under cut-off point to be approved and therefore to allow for the collection of the performance information on the rejected population, as Meng and Schmidt suggested [15]. A proxy to this approach is to collect performance information on the rejected from credit bureau using customers' overall performance across all banks. Some unpublished works using this approach appeared to show that the reject inference actually worked. In summary, the evidence on the effectiveness of reject inference is patchy and inconclusive.

It is not uncommon that modellers encounter situations where the performance of a new scorecard appears to lose significant part of its predictive power when being validated against the out-of-time sample using the Accept-Only population and as a consequence, the scorecards fail to pass the User Acceptance hurdle. Let us take a look at a real example. The performances of a new application scorecard suite compared to the existing scorecard suite are shown in Table 1 below.

Table 1 Application Scorecard Performance Example (Accept-Only Population)

Segment	Scorecard Applied	Development		Out of Time	
		Gini	KS	Gini	KS
Segment #1	Current Score	54.73%	0.43	59.24%	0.49
	New Score	57.16%	0.43	57.61%	0.44
	Uplift in Gini	**4.44%**		**-2.75%**	
Segment #2	Current Score	48.57%	0.37	49.13%	0.38
	New Score	52.94%	0.40	49.99%	0.40
	Uplift in Gini	**9.00%**		**1.75%**	
Segment #3	Current Score	34.64%	0.26	33.03%	0.25
	New Score	**39.79%**	0.30	**34.46%**	0.25
	Uplift in Gini	**14.89%**		**4.33%**	
Overall Suite	Current Score	52.87%	0.41	55.06%	0.43
	New Score	56.54%	0.43	55.15%	0.43
	Uplift in Gini	**6.94%**		**0.16%**	

It can be seen that although the uplift gained from the new scorecards appeared strong in the Accepts Only development sample, this does not hold for out-of-time. More critically, this is not a one-off phenomenon as similar patterns are observed in scorecards across the industry. It puzzled scorecard practitioners and risk managers alike. It indicates that something has been missed or overlooked in the scorecard development process, particularly in model design. It is important to understand the root causes so that it can be prevented or dealt with.

Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com

3. Investigations

Given the nature of the problem, investigation was focused on two hypotheses:

(1) effect of reject inference

(2) population shift between development and out-of-time samples

This paper is to summarise the key learning points obtained from the investigation into the Segment #3 above. But the findings can be generalised.

3.1 Reject Inference Investigation

The first line of investigation is to test the scorecard sensitivity towards Rejects sample size in the development sample. Models were built by varying the proportion of rejects included within the development sample and 4 scenarios were created as below:

- **Model A** Ratio of Accepts:Rejects is 1:0
- **Model B** Ratio of Accepts:Rejects is 1:1
- **Model C** Ratio of Accepts:Rejects is 1:2
- **Model D** Ratio of Accepts:Rejects is 1:3

These models were then applied to the Accept Only population to compare their performances (see Table 2):

Table 2 Model Performance Sensitivity to Accept/Reject Sample Ratio (Measured by model GINI using Accept Only Population)

Model	Accepts:Rejects	Development Accepts Only			Out of Time Accepts Only		
		Current Score	New Score	Uplift	Current Score	New Score	Uplift
A	1:0	34.64%	42.96%	24.02%	33.01%	34.65%	4.97%
B	1:1	34.64%	41.11%	18.68%	33.01%	35.73%	8.24%
C	1:2	34.64%	40.12%	15.82%	33.01%	35.15%	6.48%
D	1:3	34.64%	39.79%	14.87%	33.01%	34.46%	4.39%

It can be seen that

- As the % of rejects within the development sample increases, model power (GINI) among development Accepts-Only degrades gradually but steadily. However, the power reduction is of a minor nature and does not explain the problem we encountered.

- Systematic reduction in model power does not materialise across all model scenarios when measured in Out-Of-Time Accepts-Only sample. In fact, all 4 models performed similarly.

Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com

This suggested that reject inference is not the root cause and instead population shift between development and Out-Of-Time samples is more likely to have caused the problem.

3.2 Population Shift

In this section, we intend to understand what has changed between the development sample and the Out-Of-Time validation sample. To understand this, we need to differentiate three scorecards as below:

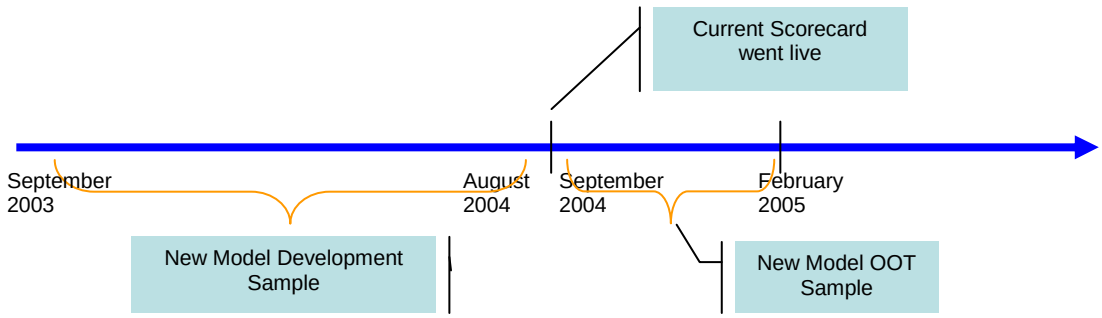
(1) Previous Scorecard: the scorecard which was replaced by the current scorecard and it is no longer in use.

(2) Current Scorecard: the existing scorecard which is currently in active use.

(3) New Scorecard: the scorecard which is newly developed to replace the existing scorecard.

Figure 1 shows the time frame used for the new model development and Out-Of-Time validation sampling vs. the Current Scorecard live date.

Figure 1 Time Frame for New Scorecard Development & Validation Sampling



Note that the Out-Of-Time sample was collected post the date Current Scorecard went live. The development sample was scored by Previous Scorecard in real life, whilst the Out-Of-Time sample was scored by the Current Scorecard. For analysis purposes, all development samples were retrospectively scored through the Current Scorecards, thus enabling a like-for-like comparison to take place.

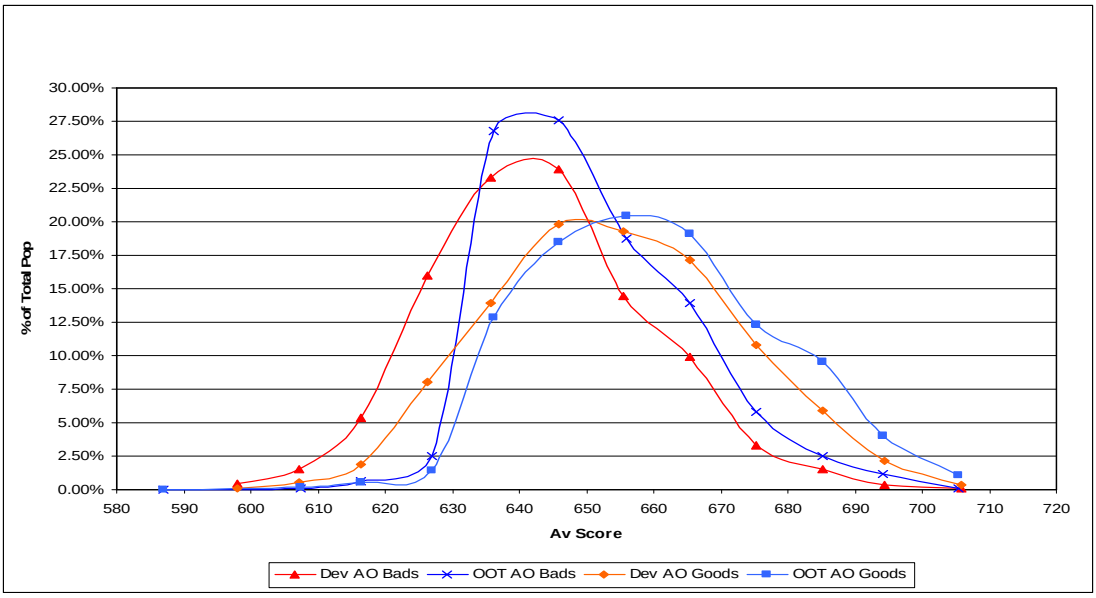
Apparently, although the development sample has been retrospectively scored through the Current Scorecard, the Previous Scorecard was actually used to make the decision on which applications to accept among Through-The-Door applicant population. Accordingly, a portion of this accepted population would have been swapped out based on the Current Score (i.e. rejected should the Current Scorecard have been used), while a portion of the rejected population would have been swapped in (i.e. accepted should the Current Scorecard have been used).

Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com

Therefore, it is logical that there will be a greater score range for the Development Accepts population when compared to the Out-Of-Time Accepts sample, where the score is assigned by using Current Scorecard.

Figure 2 Development Accept-Only vs. OOT Accept-Only Population Comparison



It is can seen from Figure 2 that there are very few accounts below the score 630 in the OOT sample, yet there are a large proportion of accounts below this score in the development sample, which are disproportionately BAD accounts.

This suggests that based on the Current Score there are a significant proportion of accounts from the development sample that would not have been accepted had we scored them in live through the Current Scorecards.

In fact, if the development Accept-Only sample's Gini is re-calculated solely on accounts with a Current Score of 630+ (the cutting point under the acquisition strategy using the Current Scorecard), the following result would be obtained.

Table 3. New Scorecard GINI for Development Accept-Only Sample (Accounts with current score 630+)

	Development		Out of time	
Population	**Current Score**	**New Score**	**Current Score**	**New Score**
Accept Only	31.8%	**37.3%**	33.7%	**34.7%**

Referring back to Table 1, the new scorecard GINI reduction from the development sample to OOT sample has now narrowed from 5.3% to 2.6% in absolute terms.

The conclusion we can make so far is as below:

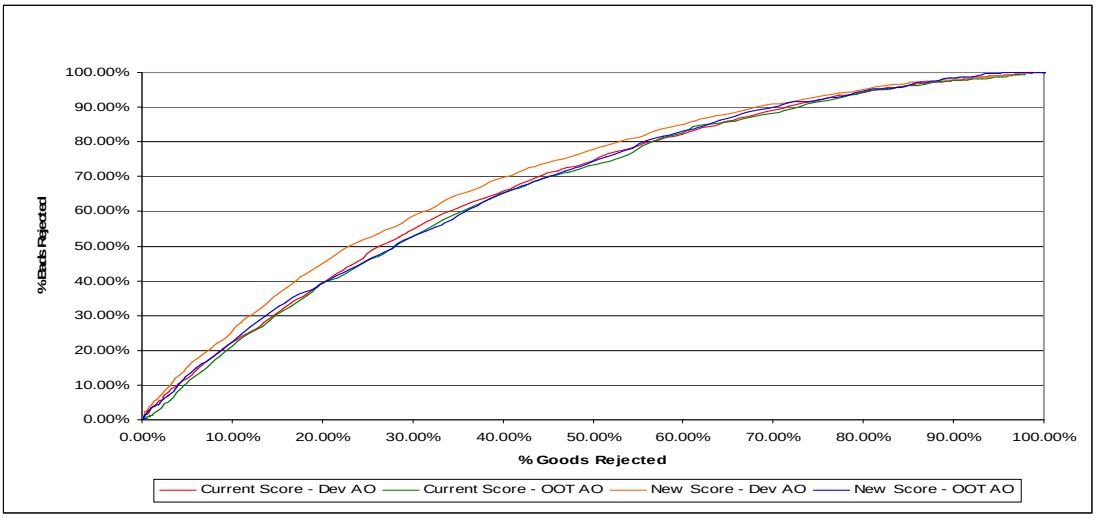
Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com

A significant portion of the New Scorecard uplift observed in development sample has already been absorbed by the Current Scorecard, the scorecard which is used to make the accept/reject decision on the OOT sample.

This will then explain why the New Scorecard's gain chart curve which appears superior in Development Accepts Only population collapses on to the Current Scorecard gains curve when measured on OOT Accepts Only. See Figure 3 below.

Figure 3. Accept-Only Population Scorecard Gains Chart Comparison



3.3 Scorecard Performance on Full Population

This paper focused on understanding the root cause for scorecard performance drop when measured against Out-Of-Time Accepts Only sample. To complete the investigation, we still want to look at how the New Scorecard performs against Through-The-Door (TTD) population on Out-Of-Time sample.

Table 4 below displays the Gini comparison based on the Through-The-Door population (performance has been inferred for rejected applications).

Table 4 Gini Comparison on Through-The-Door Population

Segment	Scorecard Applied	Dev Gini	OOT Gini
Segment #1 (Full Population)	Current Score	41.62%	44.48%
	New Score	51.89%	58.44%
	Uplift in Gini	**24.68%**	**31.38%**
Segment #2 (Full Population)	Current Score	53.18%	50.38%
	New Score	58.93%	55.07%
	Uplift in Gini	**10.81%**	**9.31%**
Segment #3 (Full Population)	Current Score	34.66%	30.74%
	New Score	43.84%	36.76%
	Uplift in Gini	**26.49%**	**19.58%**
Overall Suite (Full Population)	Current Score	49.05%	49.90%
	New Score	56.06%	58.85%
	Uplift in Gini	**14.29%**	**17.94%**

Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com

Let's still focus on the segment #3 scorecard. It can be seen that

- Both Current Scorecard and New Scorecard GINI's dropped. This may well be because of population shift. In fact, population (measured in percentage) in segment #1 were more than doubled while that in segment #3 reduced by 45%).

- New Scorecard still significantly outperforms the Current Scorecard in TTD population. Given that the New Scorecard uplift gained on development was significantly lost on OOT within Accepts Only population, most of the improvement in the TTD population actually comes from the Reject Inference samples. Swap set analysis just re-confirmed that, where BR stands for Bad Rate

Table 5. Swap Set Analysis based on Out-Of-Time Sample

	Swap ins				Swap outs				Change (%)	
	Known		Inferred		Known		Inferred		Known	Inferred
Population	Vol.	BR	Vol.	BR	Vol.	BR	Vol.	BR		
Segment #1	652	9.78%	1,383	5.48%	726	5.40%	1,170	12.12%	-44.8%	121.0%
Segment #2	711	6.42%	913	4.92%	957	8.67%	799	9.46%	35.1%	92.1%
Segment #3	473	17.97%	684	13.68%	523	17.21%	594	21.63%	-4.2%	58.1%
Suite Total	**1,836**	**10.59%**	**2,980**	**7.19%**	**2,206**	**9.62%**	**2,563**	**13.49%**	**-9.1%**	**87.6%**

4. Conclusions

This paper investigates into a practical challenge facing scorecard developers and risk managers, i.e. the performance of a newly built scorecard could deteriorate significantly when validated on Out-Of-Time samples using Accepts-Only population. Based on numerical evidences and analyses, we have reached some general conclusions. We can then present a few practical guidelines for modellers and risk managers to consider when they have to deal with the similar challenge.

4.1 General Learning Points

1. The loss of power in scorecard's GINI between Development and Out-of-Time **Accepts-Only** samples can be mainly attributed to the fact that development sample was selected prior to the date when Current Scorecard went live, whilst Out-Of-Time sample was selected after that. As a result, part of the uplift evident in the New Scorecard would have already been absorbed by the Current Scorecard.

2. Reject inference sample size appears to have some impact on development **Accepts-Only** GINI, but the impact is much less significant when assessed on Out-Of-Time Accepts Only population.

Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com

3. Performance validation against **Through-The-Door** population showed that the New Scorecard can still have significant improvement in GINI, but mainly from Reject Inference samples. Swap set analysis just re-confirmed that.

4.2 Practical Guidelines

1. Wherever possible, scorecard development sample should be selected from a period after the Current Scorecard live date.

2. Historically, banks tend to choose a long performance window between 18-24 months for application scorecards and up to 18 months for behaviour scorecard. There was good reason for doing so. But given the frequency of scorecard re-building expected nowadays (as a general guide, scorecards are re-built once every two years), one need to consider whether performance window can be reduced, particularly for application scorecard, to 12 months.

3. In some cases, however, it may not be possible to select development sample after Current Scorecard live date. For instance, the scorecard has to be replaced for various reasons, but the available data history under the Current Scorecard influence is not long enough to satisfy the desirable minimum performance window definition of the New Scorecard. Under such circumstances, one should be aware that Out-Of-Time validation GINI measured on Accepts Only population will drop (other things being equal). It is probably safe to assume that more performance reduction would be expected if the improvement of the Current Scorecard over the Previous Scorecard is bigger.

4. Risk managers need to consider how to incorporate Reject Inference based scorecard performance evidence into their scorecard implementation decision-making and strategy development, particularly when the above described circumstances occur (see Point 3), e.g. at least giving partial credit to Reject Inference sample performance.

5. Wherever possible, one should consider taking Bureau Retro samples from the Rejects as part of model development sample preparation. It would be most appropriate if a random group of rejected applicants could be identified who subsequently were successfully approved for similar financial product with a competitor within a specified period of time.

Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com

References

[1] Lewis EM (1992). An Introduction to Credit Scoring. Athena Press: San Rafaei, US.

[2] Nelson RW (1997). Credit Card Risk Management. Warren Taylor Publications.

[3] Mays E (eds) (1998). Credit Risk Scoring: Design and application. Glenlake Publishing Company Ltd.

[4] Thomas LC, Edelman DB and Crook JN (2002). Credit Scoring and Its Applications. SIAM.

[5] Desai VS, Crook JN and Overstreet GA (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research* **95**: 24-37.

[6] Desai VS, Convay DG, Crook JN and Overstreet GA (1997). Credit scoring models in the credit environment using neural networks and genetic algorithms. *IMA J Maths Appl Bus Ind* **8**: 323-346.

[7] Banasik J, Crook JN and Thomas LC (1999). Not if but when will borrowers default. *J Opl Res Soc* **50**: 1185-1190.

[8] Narain B (1992). Survival analysis and the credit granting decision. In: Thomas LC, Crook JN and Edelman DB (eds). *Credit Scoring and Credit Control*. Oxford University Press: Oxford, pp.109-121.

[9] Schedesch KB and R Stecking (2005). Support vector machine for classifying and describing credit application: detecting typical and critical regions. *J Opl Res Soc* **56**: 1082-1088.

[10] Hand DJ, McConway KJ and Stanghellini E (1997). Graphical models of applicants for credit. *IMA J Maths Appl Bus Ind* **8**: 143-155.

[11] Stanghellini E, McConway KJ and Hand DJ (1999). A discrete variable chain graph for applicant fro bank credit. *J Royal Stat Sco Ser C* **48**: 239-251.

[12] Moffatt PG (2005). Hurdle models of loan default. *J Opl Res Soc* **56**: 1063-1071.

[13] Hand, DJ and Henley WE (1993/4). Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry.* **5**: 45-55.

Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com

[14] Chen, GG (2001). *The Economic Value of Reject Inference in Credit Scoring*. Department of Management Science, University of Waterloo, Canada.

[15] Meng, CL and Schmidt P (1985). On the cost of partial observation in the bivariate probit model. *International Economic Review.* **26**: 71-85.

Correspondence: Dr. Edward X M Huang, HBOS Bank. Terra Nova House, Pier Head Street, Cardiff CF10 4PB, UK.

E-mail: edwardhuang@HBOSplc.com