Journal of Empirical Finance 17 (2010) 501–512

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Variable reduction, sample selection bias and bank retail credit scoring

Andrew Marshall [a,*], Leilei Tang [a], Alistair Milne [b]

[a] University of Strathclyde, Glasgow, G4 0NL, Scotland, UK
[b] Cass Business School, 106 Bunhill Row, London, EC1Y 8TZ, UK

ARTICLE INFO

Article history:
Received 20 April 2009
Received in revised form 3 December 2009
Accepted 9 December 2009
Available online 16 December 2009

JEL classification:
G21
C52
C53

Keywords:
Bootstrap variable selection
Credit scoring
Loan performance forecasting
Sample selection bias

ABSTRACT

This paper investigates the effect of including the customer loan approval process to the estimation of loan performance and explores the influence of sample selection bias in predicting the probability of default. The bootstrap variable reduction technique is applied to reduce the variable dimension for a large data-set drawn from a major UK retail bank. The results show a statistically significant correlation between the loan approval and performance processes. We further demonstrate an economically significant improvement in forecasting performance when taking into account sample selection bias. We conclude that financial institutions can obtain benefits by correcting for sample selection bias in their credit scoring models.

1. Introduction

The sub-prime mortgage lending crisis has shaken the financial stability of many developed countries. Many factors have caused the crisis but it has highlighted the importance of accurately assessing risks in bank retail lending. Accurate assessment of the probability of default in retail bank lending can help banks classify their customers, charge them appropriately, and improve the efficiency of credit funding. It can also result in banks having a competitive advantage over their rivals. Even a small improvement of predicting the probability of default can bring lenders substantial additional profits (see Blöchlinger and Leippold, 2006). Finally, precise risk assessment is important as the Basel Capital Accord (Basel II) requires the banking industry to meet minimum regulatory capital requirements based on the calculation of the probability of loan default, and therefore more accurate default assessment allows for more efficient utilisation of regulatory capital, a further source of competitive advantage.

The primary issue of credit scoring research has been to determine what variables significantly influence the probability of default. A second important issue in the construction of credit scoring cards is the shift from monitoring the loan performance process to the broader criteria that includes the loan approval decision process. This paper investigates the effect of including the customer loan approval process in the estimation of loan performance and explores the influence of sample selection bias in predicting the probability of default. We compare the forecasting performance of the bootstrap variable reduction procedure with single-stepwise. We expect bootstrap variable selection can reduce the likelihood of the incorrect inclusion of "noise" variables. As well as investigating further important issues in the probability of default the motivation for this paper lies in the continuing interest in sample selection bias analysis and, in particular, ongoing concerns regarding the prediction accuracy of sample selection bias.

* Corresponding author. Department of Accounting and Finance, University of Strathclyde, Glasgow, G4 0NL, Scotland, UK. Tel.: +44 141 5483894.
E-mail address: a.marshall@strath.ac.uk (A. Marshall).

0927-5398/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jempfin.2009.12.003

In this paper we use a large personal loan data-set from one of the largest UK banks to conduct a bootstrap variable reduction simulation. We quantify the sample selection bias by taking account of the correlation between the loan approval process and the subsequent loan performance process. The size of our data-set allows us to reserve a large holdout sample. Thus we can further compare the benefits from comparing the forecast performance of a loan performance process which corrects for sample selection bias against an alternative. We find that the bootstrap variable selection procedure can choose more robust explanatory variables that forecast well out-of-sample when compared with single-stepwise variable selection procedure. In particular, the variables selected by the bootstrap simulation technique can enable us to infer that the UK bank providing our data applies a credit risk minimization lending policy. We confirm that there is a statistically significant correlation between the loan granting and performance processes. We show that there is an economically as well as statistically significant improvement in forecasting performance for out-of-sample when taking into account sample selection bias. Our results are important for financial institutions who can obtain benefits by correcting for sample selection bias in their credit scoring models by including the customer loan approval process in the estimation of loan performance process.

The organization of this paper is as follows. Section 2 provides some background and a brief review of the relevant literature on retail credit scoring. Section 3 introduces the data-set we use and explains our use of a bootstrap simulation technique. In Section 4, the multi-process probit model is specified for customer approval and loan performance processes. The estimation results are presented in Section 5 where we compare the out-of-sample forecasting performance. Section 6 discusses the implications of these results and summarises our conclusions.

2. Background and literature review

The primary issue of credit scoring research has been to determine the variables that significantly influence the probability of default (see Thomas, 2000 and a recent example is Dinh and Kleimeier, 2007).[1] Typical bank retail loan databases have the salient characteristic of hundreds of variables for each customer's credit history. The high dimension of variables makes it operationally difficult to classify customers and identify the impacts of explanatory variables on estimating the probability of defaults. The relevance of explanatory variables is generally regarded as the most important consideration in the construction of credit scoring cards. Most studies into the estimation of the probability of loan default for bank retail lending have relied on rather arbitrary methods of explanatory variable selection. Such variable reduction methods mainly rely on a single-stepwise procedure or *ad-hoc* bank expert judgement system. Problems associated with a single-stepwise approach include noise in explanatory variables which bias the credit scoring cards and potentially excluding genuine explanatory variables.

A second important issue in the construction of credit scoring cards is the shift in focus from monitoring loan performance process to broader criteria that includes loan approval decision process. When a customer approaches a bank to make a loan, the bank first needs to decide whether to grant or reject the credit line request based on the customer's background and other relevant financial activity histories. If the request is accepted, then the bank can observe whether the customer in fact defaults or instead performs satisfactorily over time. Therefore there are two procedures: the credit granting process (accept or reject) and loan performance process (good or bad).[2] If a bank fits a model to predict the performance of a pool of loans it has made based on an extensive list of characteristics of the borrowers, this model will be unlikely to accurately predict the performance of a random sample of borrowers who have not yet been approved for loans by the bank. In the first instance, the fact that the bank has pre-processed the data and approved or already made the loan means that the pool of outstanding loans are not a random sample, but instead represent a selected sample relative to the population of potential borrowers. The relation between loan performance and borrower characteristics in the selected sample could differ from the performance relationships that exist in a randomly sampled pool of borrower applications. The seminal study in this vein is Boyes et al. (1989), which shows that the sample selection bias occurs if a bank credit scoring relies only on a loan performance process without considering the loan approval process. It also helps lead to the joint estimation and correlation testing between the two processes (for example, see Jacobson and Roszbach, 2003). Both of these issues are essential for the construction of a valid credit scoring card. While methods on sample selection bias are well established (Vella, 1998), there is little empirical work on robust variable reduction procedures for credit scoring card. This paper attempts to integrate the two issues (loan performance and loan approval) and compare the forecasting performance between the bootstrap variable reduction procedure and the more common single-stepwise procedure. The rationale for this is that the bootstrap variable selection can reduce the likelihood of the incorrect inclusion of "noise" variables that should not be part of the model.

3. Data description and bootstrap variable selection procedure

The unique data-set in this paper is supplied by one of the biggest commercial banks in the UK between the years of 1995 and 2003. The data are personal loans for the existing bank customers but excludes both mandatory-accept and automatic-decline customers.[3] There are three groups of characteristics on 43,634 applicants together with a summary of the subsequent

[1] Santos Silva and Murteira (2009) consider the probability of default for a particular loan that is repaid in regular instalments and develop a model to estimate conditional probability to default using data from borrowers currently paying their loans.

[2] Similarly Dionne et al. (1996) estimate jointly the default probability and two conditional truncated distributions of non-payments of good and bad loans and find that the significant variables that affect the distributions are not the same.

[3] Details on the types of personal loans such as whether the loans are secured or unsecured personal lending, revolving or repayment, etc., are not provided by the bank. Also there is no information on mandatory-accept or automatic-decline customers (decided by manually for strategic and/or legal reasons e.g. aged under 18). However, this should not lead to a biased sample for this paper (Hand and Henley, 1997).