International Journal of Forecasting 28 (2012) 224–238

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

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

Instance sampling in credit scoring: An empirical study of sample size and balancing

Sven F. Crone [1], Steven Finlay *

Lancaster University, United Kingdom

ARTICLE INFO

Keywords:
Credit scoring
Data pre-processing
Sample size
Under-sampling
Over-sampling
Balancing

ABSTRACT

To date, best practice in sampling credit applicants has been established based largely on expert opinion, which generally recommends that small samples of 1500 instances each of both goods and bads are sufficient, and that the heavily biased datasets observed should be balanced by undersampling the majority class. Consequently, the topics of sample sizes and sample balance have not been subject to either formal study in credit scoring, or empirical evaluations across different data conditions and algorithms of varying efficiency. This paper describes an empirical study of instance sampling in predicting consumer repayment behaviour, evaluating the relative accuracies of logistic regression, discriminant analysis, decision trees and neural networks on two datasets across 20 samples of increasing size and 29 rebalanced sample distributions created by gradually under- and over-sampling the goods and bads respectively. The paper makes a practical contribution to model building on credit scoring datasets, and provides evidence that using samples larger than those recommended in credit scoring practice provides a significant increase in accuracy across algorithms.

1. Introduction

The vast majority of consumer lending decisions, whether to grant credit to individuals or not, are made using automated credit scoring systems based on individuals' credit scores. Credit scores provide an estimate of whether an individual is a "good" or "bad" credit risk (i.e. a binary classification), and are generated using predictive models of the repayment behaviours of previous credit applicants whose repayment performances have been observed over a period of time (Thomas, Edelman, & Crook, 2002). A large credit granting organization will have millions of customer records and recruit hundreds of thousands of new customers each year. Although this provides a rich source of data upon which credit scoring models can be constructed, the size of the customer databases means that they often prove ineffective or inefficient (given the cost, resource and time constraints) for developing predictive models using the complete customer database. As a consequence, the standard practice has been for credit scoring models to be constructed using samples of the available data. This places particular importance on the methods applied for constructing the samples which will later be used to build accurate and reliable credit scoring models.

Despite its apparent relevance, past research in credit scoring has not systematically evaluated the effect of instance sampling. Rather than follow insights based upon empirical experiments of sample size and balance, certain recommendations expressed by industry experts have received wide acceptance within the credit scoring community and practitioner literature, driven by the understanding that customer databases in credit scoring show a high level of homogeneity between different lenders and across geographic regions. In particular, the advice of Lewis (1992) and Siddiqi (2006) is generally taken, based

* Correspondence to: Management Science, Lancaster University, Management School, Lancaster, United Kingdom. Tel.: +44 1772 798673.
E-mail addresses: s.crone@lancaster.ac.uk (S.F. Crone), steve.finlay@btinternet.com (S. Finlay).
[1] Tel.: +44 1524 5 92991; fax: +44 1524 844885.

0169-2070/$ – see front matter
doi:10.1016/j.ijforecast.2011.07.006

on their considerable experience of scorecard development. With regard to a suitable sampling strategy, both propose random undersampling in order to address class imbalances, and suggest that a sample containing 1500–2000 instances of each class (including any validation sample) should be sufficient for building robust high quality models. Given the size of empirical databases, this is equivalent to omitting large numbers of instances of both the majority class of 'goods' and the minority class of 'bads'. Although this omits potentially valuable segments of the total customer sample from the model building process, these recommendations have not been substantially challenged, either in practice or in academic research, the latter of which has focussed instead on comparing the accuracies of different predictive algorithms on even smaller and more unbalanced datasets. As a consequence, issues of sample size and balancing have been neglected within the credit scoring community as a topic of study.

Issues of constructing samples for credit scoring have only received attention in the area of reject inference, which has emphasised sampling issues relating to the selection bias introduced as a result of previous decision making in credit scoring, and the application of techniques to adjust for this bias (Banasik & Crook, 2007; Kim & Sohn, 2007; Verstraeten & Van den Poel, 2005). However, this research does not consider the more practical issues of efficient and effective sample sizes and (im-)balances. Therefore, beyond a common sense agreement that larger sample sizes are beneficial and smaller ones are more efficient, the issue of determining an efficient sample size and sample distribution (balancing) to enhance the predictive accuracy of different algorithms on the available data has not been considered. (Similarly, limited attention has been paid in credit scoring to other data preprocessing issues, such as feature selection — see Liu & Schumann, 2005; Somol, Baesens, Pudil, & Vanthienen, 2005 — or transformation — see e.g. Piramuthu, 2006 — which are deemed important but are beyond this discussion.) Considering that data and their preparation are considered to be the most crucial and time-consuming aspect of any scorecard development (Anderson, 2007), this omission is surprising and indicates a significant gap in the research.

In contrast to credit scoring, issues of sample imbalances have received a substantial amount of attention in data mining, leading to the development of novel techniques, e.g., using instance creation to balance sampling (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), frameworks for modelling rare data (Weiss, 2004), and best practices for oversampling through instance resampling (for an overview see Chawla, Japkowicz, & Kolcz, 2004), which have not been explored in the area of credit scoring. Since proven alternatives to instance sampling exist, they warrant a discussion and empirical assessment for their application to credit scoring.

In this paper, two aspects of the sampling strategy are explored in regard to their empirical impact on model performance for datasets of credit scoring structure: sample size and sample balance. Section 2 reviews the prior research, in both best practice and empirical studies, and identifies a gap in the research on instance sampling. Both the sample size and the balance are discussed, with reflections as to whether the sample size remains an issue for scorecard developers today, given the computational resources available, and investigating how random oversampling and undersampling may aid in predictive modelling. An empirical study is then described in Section 3, examining the relationship between the sample strategy and the predictive performance for two industry-supplied data sets, both larger (and more representative) than those published in research to date: one an application scoring data set, the other a behavioural scoring data set. A wide variety of sampling strategies are explored, in the form of 20 data subsets of gradually increasing size, together with 29 samples of class imbalances by gradually over- and under-sampling the number of goods and bads in each subset respectively. Having looked at both sample sizes and balancing in isolation, the final part of the paper considers the interaction between sample sizes and balancing and looks at the way in which predictive performance covaries with each of these dimensions. All of the results from the sampling strategy are assessed across four competing classification techniques, which are well established and are known to have practical applications within the financial services industry: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART) and artificial Neural Networks (NN). The empirical evaluation seems particularly relevant in light of the differences between the statistical efficiencies of the estimators with regard to the sample size and distribution, e.g. the comparatively robust Logistic Regression versus Discriminant Analysis (see, e.g. Hand & Henley, 1993). Consequently, we anticipate different sensitivity (or rather robustness) levels across different classifiers, which may explain their relative performances, beyond practical recommendations to increase sample sizes and/or balance distributions across algorithms in practice.

2. Instance selection in credit scoring

2.1. Best practices and empirical studies in sampling

The application of algorithms for credit scoring requires data in a mathematically feasible format, which is achieved through data preprocessing (DPP) in the form of data reduction, with the aim of decreasing the size of the datasets by means of instance selection and/or feature selection, and data projection, thus altering the representation of data, e.g. by the categorisation of continuous variables. In order to assess prior research on instance selection for credit scoring, best practice recommendations (from practitioners) are reviewed in contrast to the experimental setups employed in prior empirical academic studies.

In credit scoring practice, the various recommendations as to sample size concur with the original advice of Lewis (1992) and Siddiqi (2006), that 1500 instances of each class (goods, bads and indeterminates) should be sufficient to build robust, high quality models (see e.g. Anderson, 2007; Mays, 2001; McNab & Wynn, 2003, amongst others). This includes data for validation, although this requires fewer cases, perhaps a minimum of 300 of each (Mays, 2001). Anderson (2007) justifies the validity of these sample size recommendations empirically, as both Anderson and Siddiqi have worked in practice for many