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

years and the recommendations appear to be sufficiently large to reduce the effects of multicolinearity and overfitting when working with correlated variables. However, he points out that no logic has been provided for the choice of these numbers, which were determined in the 1960s when the collection of data was more costly, and that their use has continued ever since without further evaluation or challenge (Anderson, 2007), although in practice larger samples are sometimes taken where available (Hand & Henley, 1997; Thomas et al., 2002).

The considered validity of these recommendations is founded upon the understanding that customer databases in credit scoring are homogeneous across lenders and regions. Indeed, the majority of lenders ask similar questions on application forms (Finlay, 2006; Thomas et al., 2002), and use standardized industry data sources such as those supplied by credit reference agencies. Although credit reference data vary from agency to agency, the general types of consumer data supplied by credit reference agencies worldwide are broadly the same, containing a mixture of credit history, public information and geo-demographic data (Jentzsch, 2007; Miller, 2003). Consequently, datasets are homogeneous regarding the features which have predictive power, i.e. these customer characteristics. (Note that we do not consider special cases of sparse and imbalanced credit datasets, such as low default portfolios, to which these characteristics and later findings do not apply.) However, as credit scoring activities are carried out by a range of organisations, from banks and building societies to retailers, other dataset attributes will differ (Hand & Henley, 1997). The sizes of datasets, although generally considered 'large', will vary from ubiquitous data on retail credit to fewer customers for wholesale credit (Anderson, 2007). Similarly, the sample distributions of datasets, although generally biased towards the goods and with relatively few bads, will vary to reflect the different risks of the lending decision. Empirical class imbalances range from around 2 : 1 of goods to bads for some sub-prime portfolios to over 100 : 1 for high quality mortgage portfolios. It is not clear how recommendations hold across heterogeneous variations of dataset properties.

Table 1 summarizes the algorithms and data conditions of sample sizes and sample balances from a structured literature review of academic publications which have employed multiple comparisons of credit scoring algorithms and methodologies (thus eliminating a range of papers evaluating a single algorithm, or minor tuned variants).

Table 1 documents the emphasis on applying and tuning multiple classification algorithms for a given dataset sample, in contrast to evaluating the effect of instance selection in credit scoring. The review yields two conclusions:

(a) Academic studies in credit scoring have ignored possible DPP parameters relating to sample size and sample distribution. If sample size and/or sample imbalances were to have a significant impact on predictive accuracy of some algorithms, the results across various different studies in credit scoring might be impaired.

(b) The datasets used in academic studies do not reflect the empirical recommendations from credit scoring practice: the relative accuracy of algorithms is assessed across much smaller datasets than 1500 instances in each class, and datasets are left imbalanced (of the original sample distribution). This questions the representativeness of prior academic findings for practice, and their comparability across datasets with substantially different data conditions. Furthermore, this echoes similar important omissions in other areas of corporate data mining, such as direct marketing (Crone, Lessmann, & Stahlbock, 2006), and warrants a systematic empirical evaluation across data conditions.

As a further observation, most studies seem to be preoccupied with the predictive accuracy, but fail to reflect other objectives such as interpretability and resource efficiency (in both time and costs), which also determine the empirical adequacy of different algorithms in practice. Beyond accuracy, the interpretability of models — and therefore whether the model is in line with the intuition of the staff — is often of even greater importance; while speed (the speed of classification itself and the speed with which a score-card can be revised) and robustness are also relevant (see e.g. Hand & Henley, 1997). Various computational intelligence methods, such as NN and SVM, have been reported to outperform standard regression approaches by a small margin (Baesens et al., 2003), in terms of accuracy, but are not used widely due to their perceived complexity, increased resources and reduced interpretability. As a consequence, logistic regression remains the most popular method applied by practitioners working within the financial services industry (Crook, Edelman, & Thomas, 2007; Thomas, Oliver, & Hand, 2005), offering a suitable balance of accuracy, efficiency and interpretability. Discriminant analysis (DA) and Classification and Regression Trees (CART) are also popular, due to the relative ease with which models can be developed, their limited operational requirements, and particularly their interpretability (Finlay, 2008). As the DPP choices in sample size and balance may impact not only the accuracy, but also the interpretability and efficiency of the algorithms, the discussion of experimental results will need to reflect possible trade-offs between objectives while assessing the relative performance of algorithms across different data conditions. Furthermore, as the algorithms exhibit different levels of statistical efficiency, we expect changes in the relative performance of some of the algorithms (i.e., DA in contrast to the robust LR, see e.g. Hand & Henley, 1993).

2.2. Sample size

Instance sampling is a common approach in statistics and data mining, and is used both for a preliminary investigation of the data and to facilitate efficient and effective model building on large datasets, by selecting a representative subset of a population for model construction (Tan, Steinbach, & Kumar, 2006). The data sample should exhibit approximately the same properties of interest (i.e., the mean of the population, or the repayment behaviour in credit scoring) as the original set of data, such that the

Table 1
Methods and samples used in empirical studies on credit scoring.

Study	Methods[1]						Dataset and samples				
	LDA	LR	NN	KNN	CART	Other	Data sets	Good cases[2]	Bad cases[2,3]	Goods: bads	Indep. vars.
Boyle, Crook, Hamilton, and Thomas (1992)	X				X	hyb.LDA	1	662	139	4.8:1	7 to 24
Henley (1995)	X	X		X	X	PP, PR	1	6851	8203	0.8:1	16
Desai, Conway, Crook, and Overstreet (1997)[4]	X	X	X			GA	1	714	293	2.4:1	18
Arminger, Enache, and Bonne (1997)	X	X	X				1	1390	1294	1.1:1	21
West (2000)	X	X	X	X	X	KD	2	360	270	1.3:1	24
								276	345	0.8:1	14
Baesens et al. (2003)	X	X	X	X	X	QDA	8	466	200	2.3:1	20
						BC		455	205	2.2:1	14
						SVM		1056	264	4.0:1	19
						LP		2376	264	9.0:1	19
								1388	694	2.0:1	33
								3555	1438	2.5:1	33
								4680	1560	3.0:1	16
								6240	1560	4.0:1	16
Ong, Huang, and Gwo-Hshiung (2005)	X	X	X			GP, RS	2	246	306	0.8:1	26
								560	240	2.3:1	31

[1] BC = Bayes Classifiers, CART = Classification and Regression Trees, GA = Genetic Algorithm, GP = Genetic Programming, KD = Kernel Density, KNN = K-Nearest Neighbour, LDA = Linear Discriminant Analysis, LP = Linear Programming, LR = Logistic Regression, NN = Neural Networks, QDA = Quadratic Discriminant Analysis, PP = Projection Pursuit, PR = Poisson Regression, RS = Rough sets, SVM = Support Vector Machines.
[2] In some studies, the number of goods/bads used for estimating the model parameters is not given. In these cases, the number of goods/bads has been inferred from information provided about the total sample size, the proportion of goods and bads and the development/validation methodology applied.
[3] This is the number of variables used for parameter estimation after pre-processing.
[4] Three data sets from different credit unions were used, and the models were estimated using both the individual data sets and a combined data set. The figures quoted are for the combined data set.

discriminatory power of a model built on a sample is comparable to that of one built on the full dataset. Larger sample sizes increase the probability that a sample will be representative of the population, and therefore ensure similar predictive accuracy, but also eliminate many of the advantages of sampling, such as reduced computation times and data acquisition costs. In smaller samples, the patterns contained in the data may be missed or erroneous patterns may be detected, thus enhancing efficiency at the cost of limiting accuracy. Therefore, determining an efficient and effective sample size requires a methodical approach, given the properties of the datasets, in order to balance the trade-off between accuracy and resource efficiency, assuming that the interpretability of the models is not impacted.

Although empirical sample sizes in credit scoring will vary by market size, market share and credit application, most data sets are large. In the UK, Barclaycard has over 11 million credit card customers, and recruited more than 1 million new card customers in 2008 (Barclay's, 2008). In the US, several organizations, such as Capital One, Bank of America and Citigroup, have consumer credit portfolios containing tens of millions of customer accounts (Evans & Schmalensee, 2005). Samples are conventionally built using a stratified random sample (without replacement) on the target variable, either drawing an equal number of goods and bads in proportion to the size of that group in the population (unbalanced), or using equal numbers of instances of each class (balanced). Consequently, the limiting factor for sample sizes is often the number of bads, with few organizations having more than a few thousand, or at most tens of thousands, of bad cases (Anderson, 2007).

From a theoretical perspective, at first sight, the issue of larger sample sizes might appear to have been resolved in recent years due to the increased computational resources provided by a modern PC. Credit scoring models using popular approaches such as LR, LDA and CART can be developed within a few hours using large samples of observations. Since using the entire available dataset, rather than sampling from it, has now become a feasible course of action for most consumer credit portfolios, it is valid to ask: is the issue of the sample size still relevant for scorecard development? In practical situations, the sample size does still remain an important issue for a range of reasons. First, there are often costs associated with acquiring data from a credit reference agency, resulting in a trade-off between the increase in accuracy obtained from using larger samples and the marginal cost of acquiring additional data. Also, a model developer may rebuild a model many times, possibly even dozens of times, to ensure that the model meets the business requirements that are often imposed on such models, or to evaluate the effect of different DPPs or different meta-parameters of algorithms, in order to enhance performance. This means that even small reductions in the time required to estimate the parameters of a single model on a sample may result in large and significant reductions in the project time/cost when many iterations of model development occur. In contrast, when sub-population models are considered, larger samples may be required. It may be relatively easy to construct a sub-population model and confirm that it generates unbiased estimates, but if the sample upon which it has been developed is too small, then a locally constructed sub-population model may not perform as effectively as a model developed on a larger, more general population, despite the statistical efficiency of the estimator. In the case of population drift, where applicant populations and distributions evolve over time due to

changes in the competitive and economic environment (Hand & Henley, 1997), not all records of past applicants are representative of current/future behaviour, and hence sampling is required. This has been particularly apparent in the recent credit crunch, where new models have needed to be constructed for the novel economic circumstances (Hand, 2009a,b), thus limiting the ability to use all data and raising the question as to what the minimum (or near optimal) sample size for a renewed scorecard development would be. Consequently, larger samples are not always desirable. Rather, the trade-off between accuracy and computational costs must be considered in order to derive resource efficient and effective decisions with respect to the sample size.

Furthermore, some algorithms are expected to perform better on larger samples of data, whilst others are more efficient at utilizing a given training sample when estimating parameters. For example, NN and SVM generally outperform LR when applied to credit scoring problems (Crook et al., 2007), but when the sample sizes are small, LR may generate better performing models due to the smaller number of parameters requiring estimation. In contrast, when datasets get very large, NN are considered to benefit from the additional data, while the performance of SVM would suffer. This implies that the sample size should be considered alongside other features when deciding upon the algorithm, and may explain previous inconsistent results on the relative accuracy of the same methods across credit scoring studies for different sample sizes.

2.3. Sample distribution (balancing)

For real-world credit scoring datasets, the target variable is predominantly imbalanced, with the majority of instances composed of normal examples (“goods”) and only a small percentage of abnormal or interesting examples (“bads”). A dataset is said to be unbalanced if the number of instances in each category of the target variable is not (approximately) equal, which is the case across most applications in both credit scoring and data mining.

The importance of reflecting the imbalance between the majority and minority classes in modelling is not primarily an algorithmic one, but is often derived from the underlying decision context and the costs associated with it. In many applications, the costs of type I and type II errors is dramatically asymmetrical, making an invalid prediction of the minority class more costly than an accurate prediction of the majority class. However, traditional classification algorithms — driven by the objective of minimising some loss of error function across two different segments of a population — typically have a bias towards the majority class, which provides more error signals. Therefore, the underlying problem requires either the development of distribution insensitive algorithms or an artificial rebalancing of the datasets through sampling.

Problems of data driven model building with imbalanced classes are not uncommon in other domains of corporate data mining, such as response rates in direct marketing, and are ubiquitous in classification tasks across a range of disciplines (see e.g. the special issue by Chawla et al., 2004). In instance sampling, random over- and undersampling methodologies have received particular attention (Weiss & Provost, 2003). In undersampling, instances of the minority and majority classes are selected randomly in order to achieve a balanced stratified sample with equal class distributions, often using all instances of the minority class and only a sub-set of the majority class, or undersampling both classes for even smaller subsets with equal class sizes. Alternatively, in oversampling, the cases of the under-represented class are replicated a number of times, so that the class distributions are more equal. Note that inconsistencies in this terminology are frequent, and also arise in credit scoring (e.g. Anderson, 2007, mistakenly refers to oversampling, but essentially describes simple undersampling by removing instances of the majority class).

Under- and over-sampling generally lead to models with a enhanced discriminatory power, but both random oversampling and random undersampling methods have their shortcomings: random undersampling can discard potentially important cases from the majority class of the sample (the goods), thus impairing an algorithm’s ability to learn the decision boundary; while random oversampling duplicates records and can lead to the overfitting of similar instances. (Note that under- and over-sampling are only conducted on the training data used for model development, while the original class distributions are retained for the out-of-sample test data.) Therefore, undersampling tends to overestimate the probability of cases belonging to the minority class, while oversampling tends to underestimate the likelihood of observations belonging to the minority class (Weiss, 2004). As both over- and under-sampling can potentially reduce the accuracy in generalisation for unseen data, a number of studies have compared variants of over- and under-sampling, and have presented (often conflicting) viewpoints on the accuracy gains derived from oversampling versus undersampling (Chawla, 2003; Drummond & Holte, 2003; Maloof, 2003; Prati, Batista, & Monard, 2004), indicating that the results are not universal and depend on the dataset properties and the application domain. However, as datasets in credit scoring have similar properties across lenders, the findings on over- vs. under-sampling are expected to be more representative across databases.

Reflecting on best practices and empirical studies (see Section 2.1), credit scoring practices actively recommend and exclusively employ undersampling, while academic studies have predominantly used the natural distribution of the imbalanced classes (see Table 1). Both have ignored the various oversampling approaches developed in data mining, and the evidence of the impaired accuracy caused by removing potentially valuable instances from the sample through undersampling. More sophisticated approaches to under- and over-sampling have been developed, e.g. selectively undersampling unimportant instances (Laurikkala, 2002) or creating synthetic examples in oversampling (Chawla et al., 2002), in addition to other alternatives such as cost sensitive learning. However, in the absence of an evaluation of even simple approaches to imbalanced instance sampling in credit scoring, these are omitted for the benefit of a systematic evaluation of different intensities of over- and

under-sampling. It should be noted that we later assess under- and over-sampling on empirical datasets which are subject to an inherent sample selection bias towards applicants who were previously considered creditworthy. Possible remedies of reject inference are ignored in this analysis. However, as instance sampling techniques of over- and under-sampling merely provide different error signals from the information contained within the original sample, and do not augment the sample for missing parts of the population, we suspect the effect of instance sampling techniques to be complementary to choices of reject inference.

Moreover, the error signals derived from different numbers of goods and bads may shift the decision surface in feature space for those methods estimating decision boundaries using fundamentally different approaches to classifier design, depending on their statistical efficiency. LR estimates the probability (P) that an applicant with a particular vector x of characteristic levels is good directly, with $P(g|\bullet)$, while the LDA estimates and the probability density function of the good-risk applicants will be denoted by $p(\bullet|g)$ and $p(\bullet|b)$ for bads respectively, and $P(g|\bullet)$ is then derived (see Hand & Henley, 1993, for a more elaborate discussion). Algorithms such as NN offer additional degrees of freedom in model building, beyond those of LR, which may yield different levels of statistical efficiency. For example, a NN may consider the prediction of goods directly by employing a single output node to estimate $P(g|\bullet)$ (essentially modelling a conventional LR with multiple latent variables, depending on the number of hidden nodes), by using two independent output nodes to assess $p(\bullet|g)$ and $p(\bullet|b)$ to derive $P(g|\bullet)$, as in LDA, or by using combinations by linking multiple output nodes using the softmax function, which pose undefined statistical properties and efficiency. Should these meta-parameter choices impact the estimator efficiency, as well as increasing the number of parameters to be estimated through latent variables, different practical recommendations for an effective dataset size and balance may be the result.

The effect of balancing on estimators of different statistical efficiency should be assessed separately to the effect of sample sizes, or the joint effect of the sample distribution and sample size. This assessment will be problematic for strong undersampling, since, by creating very small samples, the parameters may deviate from the population value somewhat, due to the inherent variance in smaller samples, despite ensuring random sampling throughout all of the experiments. To reflect this, our experiments will also include small sample sizes, not to compare classifiers of different statistical efficiency across these small samples, but rather to replicate the inconsistent findings of many academic studies. It is anticipated that small sample sizes should result in inconsistencies in relative classifier accuracy levels, caused predominantly by experimental biases introduced through the arbitrarily chosen small sample size, but not in the classifiers' capabilities (i.e., non-linearity etc.) or the sample distribution of the data. Such findings would confirm the findings of previous studies, and hence add to the reliability of our findings, and place our assessment of the sample size and balance in the context of the existing research.

Furthermore, it should be noted that over- and under-sampling will impact not only the predictive accuracy, depending on the statistical efficiency, but also the resource efficiency in model construction and application. Balancing has an impact on the total sample size by omitting or replicating good and/or bad instances, thereby decreasing or increasing the total number of instances in the dataset, which impacts the time taken for model parameterisation (although this seems less important than improving the accuracy, as the time taken to apply an estimated model will remain unchanged).

3. Experimental design

3.1. Datasets

Two datasets, both of which are substantially larger than those used in empirical studies to date (see Table 1), were used in the study, taken from the two prominent sub-areas of credit and behavioural scoring.

The first dataset (dataset A) was supplied by Experian UK, and contained details of credit applications made between April and June 2002. Performance information was attached 12 months after the application date. The Experian-provided delinquency status was used to generate a 1/0 target variable for modelling purposes (good = 1, bad = 0). Accounts which were up-to-date, or no more than one month in arrears, and which had not been seriously delinquent within the last 6 months (three months or more in arrears) were classified as good. Those that were currently three or more months in arrears, or had been three months in arrears at any time within the last 6 months, were classified as bad. This is consistent with the good/bad definitions commonly reported in the literature as being applied by practitioners, based on bads being three or more cycles delinquent and goods as up-to-date or no more than one cycle delinquent (Hand & Henley, 1997; Lewis, 1992; McNab & Wynn, 2003). After the removal of outliers and indeterminates, the sample contained 88,789 observations, of which 75,528 were classified as good and 13,261 as bad. 39 independent variables were available in set A. The independent variables included common application form characteristics such as age, residential status and income, as well as UK credit reference data, including the number, value and time since the most recent CCJ/bankruptcy, current and historical account performance, recent credit searches, and Electoral Roll and MOSAIC postcode level classifiers.

The second dataset, dataset B, was a behavioural scoring data set from a mail order catalogue retailer providing revolving credit. Performance data were attached as at 12 months after the sample date. The good/bad definition provided by the data provider was similar to that for set A. Goods were defined as being no more than one month in arrears, bads as being three or more months in arrears at the outcome point. After exclusions such as newly opened accounts (less than 3 months old), dormant accounts (maximum balance on the account within the last 3 months = £0), accounts already in a serious delinquency status (currently 2+ payments in arrears), and those classified as indeterminate (neither good nor bad), the

sample contained 120,508 goods and 18,098 bads. Dataset *B* contained 55 independent variables, examples of which were current and historic statement balances, current and historic arrears status, payment to balance ratios, and so on.

3.2. Sample size

The first part of the study looked at the effects of increasing the sample size on the predictive performance. For the purpose of the study, and to ensure valid and reliable estimates of the experimental results despite some small sample sizes, we employed *k*-fold random cross-validation across all experiments, essentially replicating each random sample $k = 50$ times (i.e., resampling). For each of the two datasets, a set of subsamples of different size were constructed using the following procedure:

Step 1. The population of *N* observations, comprising *G* goods and *B* bads ($N = G + B$) was segmented into *k* sections of equal size, with *p* percentiles within each fold. Stratified random sampling was applied, with the goods and bads sampled independently to ensure that the class priors in each section and percentile matched that of the distribution in the population.

Step 2. A *k*-fold development/validation methodology was applied to construct *k* models for each cumulative *p* percentage of the population. The number of observations used to construct each model, N_p, was therefore equal to $N * [p * (k - 1)/k]/100$. N_{pg} andN_{pb} are the number of goods and bads used to construct each model, such that $N_p = N_{pg} + N_{pb}$. For each model, all N/k observations in the validation section were used to evaluate the model performance.

Values of *p* ranging from 5% to 100% were considered in increments of 5%, in order to evaluate any consistent and gradual effects of the sample size variation on accuracy, leading to 20 different sample sizes. For a relationship between the sample size and accuracy, we would expect consistent, statistically significant results of increasing accuracy (i.e., a monotonically increasing trend of improving performance for the results to be considered reliable) beyond the recommended "best practice" sample size of 1500–2000 bads. The number of percentiles was chosen under the constraint of the available observations and the number of variables, so that all of the variables would still contain significant numbers of observations and allow stable parameter estimates when the sample sizes were small (a minimum of 250 bads and 500 goods for $p = 5$). To comply with what is reported to be standard practice within the credit scoring community, balanced data sets were used, with the goods being randomly under-sampled (excluded) from each section for model development, so that the number of goods and bads was the same.

3.3. Balancing sample distributions

The second part of the study considered balancing. In data mining in general, studies to date have been conducted using undersampling on the original distribution of the population, or oversampling on algorithms of varying statistical efficiency. However, this does not allow for inference on the possible systematic and continuous effects of decreasing the number of instances from the majority class (undersampling) or increasing the number of the minority class (oversampling) during stratified sampling. Therefore, for this part of the experiment multiple random samples of gradually increasing class imbalances were created from the full data set (i.e. $p = 100$), with a different balancing applied to each sample. In total, 29 different balancings were applied. For descriptive purposes we refer to each balancing using the notation B_x. The 29 different balancings were chosen on the basis of expert opinion, taking into account computation requirements and the need to obtain a reasonable number of examples across the range.

To create each undersampled data set, observations were randomly excluded from the majority class (the goods) to achieve the desired number of cases. B_{12} represents the original class imbalanced sample. Samples B_1–B_{11} were randomly under-sampled to an increasing degree of class imbalance, with B_3 representing standard undersampling, with the goods sampled down to equal the number of bads, and B_2 undersampling the goods beyond the number of bads (i.e., fewer goods than bads). B_{13}–B_{22} were randomly oversampled with increasing class imbalances, with sample B_{22} representing standard oversampling, with the bads re-sampled so that the number of goods and bads was equal. For B_{23}–B_{29}, the oversampling was extended further, so that the samples contained more bads than goods.

This creates a continuous, gradually increasing imbalance from extreme undersampling to extreme oversampling, spanning most of the sampling balances employed in data mining, while allowing us to observe possible effects from a smooth transition of accuracy due to sample imbalances.

To create the oversampled data sets, each member of the minority class (the bads) was sampled $\mathrm{INT}(N_{pv}/N_{pb})$ times, where N_{pv} is the desired number of bads in the sample (thus, for standard over-sampling, where the number of bads is equal to the number of goods, $N_{pv} = N_{pg}$). An additional $(N_{pv} - \mathrm{INT}(N_{pv}/N_{pb}))$ bads were then randomly sampled without replacement, so that the sample contained the desired number of observations (N_{pv}). The *k*-fold development/validation methodology described in Section 3.1 was adopted, with the observations assigned to the same 50 sections. Note that no balancing was ever applied to the test section; i.e., the class priors within the test section were always the same as those in the unbalanced parent population from which it was sampled.

The third and final part of the analysis considered the sample size and balancing in combination. The balancing experiments described previously were repeated for values of *p* ranging from 5% to 100% in increments of 5. In theory, this allows a 3-D surface to be plotted to show how the sample size, balancing and performance co-vary, and makes it possible to consider trade-offs between the sample size and balancing. It is noted that part 3 represents a superset of experiments, containing all of those described in parts 1 and 2, as well as many others. We have taken this approach, building up the results in stages, to increase the readability of the paper.

3.4. Methods, data pre-processing and variable selection

The methods were chosen to represent those established in credit scoring, including LR, LDA and CART, as well as NN, a frequently evaluated contender which has shown an enhanced accuracy in fraud detection and other back-end decision processes (where limited explicability is required, see e.g. Hand, 2005), but which has failed to prove its worth in credit scoring so far. As the evaluation of different modelling techniques is not of primary interest in this study, recently developed methods such as SVM are not assessed.

The experiments were repeated for LR, LDA, CART and NN. For the CART and NN models, the development sample was further split 80/20 for training/validation using stratified random sampling. For CART, a large tree was initially grown using the training sample, then pruning was applied using the 20% validation sample, as advocated by Quinlan (1992). Binary splits were employed, based on maximum entropy. For NN, a MLP architecture with a single hidden layer was adopted. Preliminary experiments were performed in order to determine the number of hidden units for the hidden layer using the smallest available sample size (i.e. $p = 5$), to ensure that overfitting did not result for small sample sizes. $T - 1$ exploratory models were created using $2, 3, \ldots, T$ hidden units, where T was equal to the number of units in the input layer. The number of hidden units was then chosen, based on the model performance on the 20% test sample. Given the size and dimensions of the datasets involved and the number of experiments performed, we employed a quasi-newton algorithm with a maximum of 100 training iterations, in order to allow the experiments to be completed in a realistic period of time.

The most widely adopted approach to pre-processing credit scoring data sets is to categorize the data using dummy variables (Hand & Henley, 1997), which generally provides a good linear approximation of the non-linear features of the data (Fox, 2000). Continuous variables such as income and age are binned into a number of discrete categories and a dummy variable is used to represent each bin. Hand (2005) suggests that between 3 and 6 dummies should be sufficient in most cases, although a greater number of dummies may be defined if a sufficient volume of data is available. For datasets A and B, the independent variables were a mixture of categorical, continuous and semi-continuous variables, which were coded as dummy variables for LDA, LR, NN. All of the dummy variables in each dataset contained in excess of 500 good and 250 bad cases, and more than 1000 observations in total (for $p = 100$). For CART, preliminary experiments showed that the performance based on dummy variables was extremely poor, and a better performance resulted from creating an ordinal range using the dummy variable definitions. This ordinal categorization was therefore used for CART. We note that the results of this particular data preprocessing strategy may be biased against some of the nonlinear algorithms (Crone et al., 2006), but it was chosen due to its prevalence in credit scoring practice and academic studies. To allow the experiments to be replicated, additional details on the data preprocessing and method parameterisation can be obtained from the authors upon request.

3.5. Performance evaluation

For measuring the model accuracy, a precise estimate of the likelihood of class membership may serve as a valid objective of parameterisation; however, it is of secondary importance to a model's ability to accurately discriminate between the two classes of interest (Thomas, Banasik, & Crook, 2001). As a consequence, measures of group separation, such as the area under the ROC curve (AUC), the GINI coefficient and the KS statistic, are used widely for assessing model performance, especially in situations where the use of the model is uncertain prior to model development, or where multiple cut-offs are applied at different points in the score distribution. Performance measures must also be insensitive to the class distribution, given the data properties of credit scoring (i.e., simple classification rates may not be applied). A popular metric in data mining, the AUC, provides a single valued performance measure [0; 1] which assesses the tradeoff between hits and false alarms, where random variables score 0.5. To employ a performance measure which is more common in the practice of the retail banking sector, we assess the model performance using the related GINI coefficient, calculated using the brown formula (Trapezium rule):

$$\text{GINI} = 1 - \sum_{i=2}^{n} [G(i) + G(i-1)]\,[B(i) - B(i-1)],$$

where S is the ranked model score and $G(S)$ and $B(S)$ are the cumulative proportion of good and bad cases, respectively, scoring $\leq S$ for all S. GINI is an equivalent transformation of the AUC (Hand & Henley, 1997), measuring twice the area between the ROC-curve and the diagonal (with $\text{AUC} = (\text{Gini} + 1)/2$), to assess the true positive rate against the false positive rate. GINI measures the discriminatory power over all possible choices of threshold (rather than the accuracy of probability estimates of class membership), which satisfies the unconditional problem of an unknown threshold or cost ratio in which GINI is considered advantageous (see, e.g., the third scenario of Hand, 2005), and which adequately reflects our empirical modeling objective. Furthermore, it allows the results to be compared directly with other studies, including applications in retail banking where practitioners regularly employ GINI, which is considered to be equally important. Therefore, despite recent criticism (see e.g. Hand, 2005, 2009a,b), the limited theoretical weaknesses of GINI seem to be outweighed by its advantages in interpretability, both by practitioners and across other studies.

4. Experimental results

4.1. Effect of sample size

The first stage of the analysis considers the predictive accuracy of methods constructed using different sample sizes for equally distributed classes (in the training data) using undersampling. The results of the sample size experiments are presented in Table 2.

Table 2 shows both the comparative level of accuracy between methods and changes in the accuracy for

Table 2
Absolute GINI by sample size for datasets *A* and *B*.

p (%)	Dataset *A*					Dataset *B*				
	# Goods/bads	LDA	LR	CART	NN	# Goods/bads	LDA	LR	CART	NN
5	663	0.704**	0.702**	0.572**	0.660**	904	0.610**	0.604**	0.536**	0.572**
10	1 326	0.721**	0.721**	0.605**	0.692**	1 809	0.635**	0.633**	0.542**	0.600**
15	1 989	0.727**	0.730**	0.634**	0.701**	2 714	0.641**	0.640**	0.548**	0.605**
20	2 652	0.729**	0.733**	0.638**	0.707**	3 619	0.646**	0.645**	0.556**	0.614**
25	3 315	0.730**	0.733**	0.634**	0.711**	4 524	0.649**	0.648**	0.567**	0.624**
30	3 978	0.731**	0.735*	0.637**	0.717**	5 429	0.649**	0.649**	0.574**	0.624**
35	4 641	0.732*	0.736*	0.638**	0.722**	6 334	0.650**	0.650**	0.572**	0.628**
40	5 304	0.733	0.736*	0.644**	0.723**	7 239	0.651**	0.651**	0.572**	0.633**
45	5 967	0.733	0.737	0.648**	0.725**	8 144	0.652**	0.652**	0.577**	0.635**
50	6 630	0.733	0.737	0.656**	0.726**	9 049	0.652**	0.652**	0.581**	0.637**
55	7 293	0.733	0.737	0.658**	0.727**	9 953	0.652**	0.653**	0.576**	0.636**
60	7 956	0.733	0.737	0.654**	0.729**	10 858	0.653**	0.653**	0.577**	0.637**
65	8 619	0.733	0.737	0.659**	0.730**	11 763	0.653**	0.654**	0.578**	0.644**
70	9 282	0.733	0.738	0.658**	0.730**	12 668	0.654**	0.655**	0.578**	0.644**
75	9 945	0.733	0.738	0.656**	0.731*	13 573	0.654**	0.655**	0.580**	0.645**
80	10 608	0.734	0.738	0.659**	0.731*	14 478	0.654*	0.655**	0.583**	0.646**
85	11 271	0.734	0.738	0.663*	0.731*	15 383	0.655*	0.656*	0.579**	0.648**
90	11 934	0.734	0.738	0.663	0.732	16 288	0.655	0.656	0.581**	0.649**
95	12 597	0.734	0.738	0.664	0.732	17 193	0.655	0.656	0.582**	0.650
100	13 261	0.734	0.738	0.664	0.732	18 098	0.655	0.656	0.588	0.651

* Indicates that the performance is significantly different from $p = 100\%$ at the 95% level of significance.
** Indicates that the performance is significantly different from $p = 100\%$ at the 99% level of significance.

each individual method as the sample size increases. Table 2 also shows the results from paired *t*-tests for determining whether there is a statistically significant difference in performance between $p = 100$, the largest possible sample available, and models constructed using samples containing only $p = x\%$ [goods/bads] ($5 \leq x \leq 100$). Observing the monotonically increasing significance of the results, the paired *t*-test is considered a valid proxy for more comprehensive non-parametric tests of repeated measures. It therefore provides a plausible assessment of the asymptotic relative efficiencies of different classification algorithms, as indicated earlier, with LR and LDA already approaching this level at 5000 bads, while NN require more than double the number of instances (while still not achieving the accuracy of LR).

Table 2 documents two original findings. First, all of the methods show monotonic increases in their predictive accuracies (with minor fluctuations in accuracy for CART), which might be considered unsurprising given the common practical understanding that more data is better. However, the accuracy increases well beyond the recommended "best practice" sample size of 1500–2000 instances of each class. For logistic regression and LDA, around 5000 samples of 'bads' are required before the performance is statistically indistinguishable from that resulting from a sample size of $p = 100$ for dataset *A*, while around 15,000 cases are required for dataset *B*. This results in significantly larger (balanced) datasets of 10,000 and 30,000 instances altogether, respectively, which far exceeds both the recommendations of practice and the experimental design of academic studies. Equally, CART and NN require larger samples before their accuracy asymptotically approaches a maxmimum value, but again datasets of a larger magnitude yield further performance improvements. It is important to note that these tests of significance between samples of size $p = x$ and $p = 100$ should only be considered as lower bounds on the optimal sample size, because the study has been limited by the number of observations in the data set, rather than the theoretical maximum possible number of observations (i.e. $N = \infty$). Another factor is the fact that the numbers of observations within each coarse classed interval were chosen so that when the sample sizes were small (e.g. $p = 5$), all of the variables would still contain a sufficient number of observations of each class to allow valid parameter estimation. In real world modelling situations, a larger number of dummy variable categories could be defined when large samples are used, which could be expected to result in an improvement in the performance of the resulting models.

The effects of an increased sample size on the algorithm performance are also illustrated in Fig. 1, which shows the relative increase in accuracy of each method, indexed relative to the results that are obtained from using the industry best practice recommendation of 1500 instances, obtained via undersampling (=100%), for both dataset *A* (Fig. 1(a)) and dataset *B* (Fig. 1(b)).

Note that Fig. 1 provides the relative improvements for each of the methods in isolation, and does not compare the performances of the methods. Fig. 1 shows similar patterns for models developed using datasets *A* and *B*, indicating a similar trend in performance with an increasing sample size. Increasing the sample sizes from 1500 to the maximum possible, improved the performance by 1.78% for LR, 1.40% for LDA, 5.11% for NN and 7.11% for CART on dataset *A*. On dataset *B*, the improvements were even more substantial, significantly increasing the performance by 3.14% for LDA, 3.72% for LR, 8.41% for NN and 8.48% for CART (although it should be noted that the absolute performance of CART was consistently worse than that of the other methods for both data sets, following the trend seen in Table 2). As statistically significant improvements are also feasible from simply increasing the sample size for the well explored and comparatively

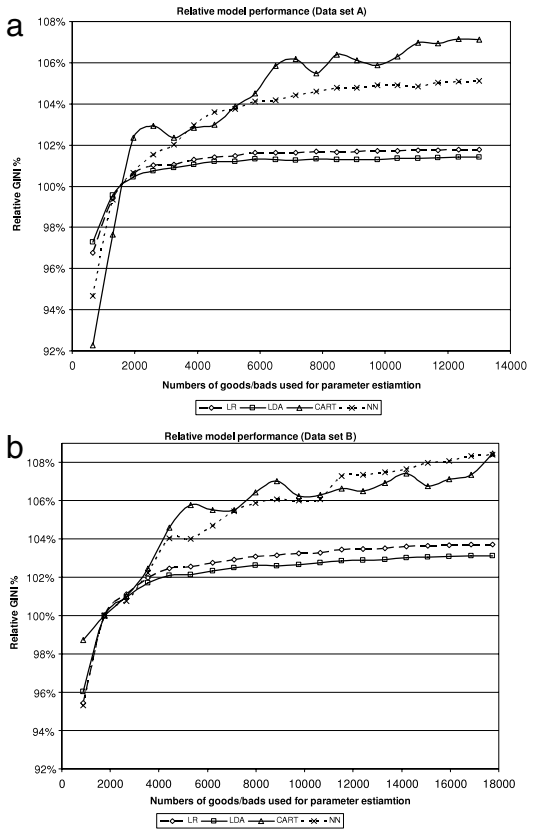

Fig. 1. Relative model performance by sample size for data sets *A* (a) and *B* (b).

efficient estimator LR, well beyond its best practices, these findings provide novel and significant advancements for credit scoring practice. All increases may be considered to be substantial, considering the flat maximum effect (Lovie & Lovie, 1986) and the costs associated with increasing the scorecard accuracy by fractions of a percentage point. Furthermore, the differences due to the sample size are substantial, considering the improvements in performance attributed to algorithm choices and tuning in the literature (see, e.g., Table 1).

The issue of an effective and efficient sample size, and of an asymptotic relative efficiency for each algorithm, is also visible in the relative performance graphs. The relative unit increase in performance of LR and LDA reduces steadily as N_{pb} rises above 2000, plateauing at around $N_{pb} = 5000$ for dataset *A* and 12,000 for dataset *B*, indicating little need for larger datasets to be collected. However, for NN and CART it would appear that the performance has not plateaued by N_{100b}, and therefore, the absolute performance might improve if larger samples were available.

The second prominent feature of Table 2 is that the relative performances of the different methods vary with the sample size, particularly for small samples. One concern which might be raised from the experimental design of the sample size is the possibility of over-fitting for small values of N_{pb}, when the ratio of observations to independent variables is low. If the 1 : 10 rule quoted by Harrell for logistic regression is taken as a guide (Harrell, Lee, & Mark, 1996), then this suggests that there is a risk of over-fitting where $N_{pb} \leq 810$ ($p \leq 6$; i.e. the first row in Table 3 and the first data point in Fig. 1). However, the region where $N_{pb} \leq 810$ is not the area of greatest interest. Also, because of the preliminary variable selection procedure, variables have only been included in the models where there is a high degree of certainty that a relationship between the dependent and independent variables exists. It is also true that the ratio of events to variables tends to be a less important factor for large samples containing hundreds of cases of each class than for smaller samples (Steyerberg, Eijkemans, Harrell, Habbema, & Dik, 2000). Therefore, we think it unlikely that over-fitting has occurred.

We conclude that increasing the sample size beyond current best practices increases the accuracy significantly for all forecasting methods considered, despite the possible negative implications for resource efficiency in model building. Moreover, the individual methods show different sensitivities to the sample size, which allows us to infer that statistical efficiency may provide one explanation for the inconsistent results on the relative performances of different classification methods on small credit scoring datasets of varying sizes, allowing LDA, LR or possibly NN to outperform other methods, depending purely on the (un-)availability of data.

4.2. Effect of balancing

The second set of analyses reviews the effect on the predictive accuracy of balancing the distribution of the target variable. Table 3 provides the results for different sample distributions B_n using all available data ($p = 100$), indicating the joint effect of changing both the sampling proportions for each of the classes and the sample sizes as a result of rebalancing.

Fig. 2 provides a graphical representation of the results.

In examining the results from Table 3 and Fig. 2, we shall begin with the performance of logistic regression. Logistic regression is remarkably robust to balancing, yielding >99.7% of the maximum performance for both data sets, regardless of the balancing strategy applied. For both data sets, undersampling leads to worse performances than either the unbalanced data set (B_{12}) or oversampling (B_{22}), and using the unbalanced data gives slightly worse performances than oversampling. However, none of these differences are statistically significant. LDA displays a greater sensitivity, with its performance falling to just under 99.4% for dataset A and 98% for dataset *B*. For both datasets, the worst performances for LDA are when B_{12} is applied, and these differences are significant at the 99% significance level. CART is by far the most sensitive technique, with a maximum performance of 95% for dataset *A* and 84% for dataset *B*. NN also shows some sensitivity to balancing, with a performance which worsens, the greater the degree of undersampling applied.

Another feature displayed in Fig. 2, and arguably the most interesting one, is that the maximum performance does not always occur at the traditional over-sampling point (B_{22}). For dataset *A*, the optimal balancing is at B_{21}, B_{17}, B_{18} and B_{16} for LR, LDA, CART and NN respectively.

Table 3
Absolute GINI by sample distribution for datasets *A* and *B*.

B_n	Dataset *A*						Dataset *B*					
	Goods	Bads	LDA	LR	CART	NN	Goods	Bads	LDA	LR	CART	NN
	807	13,261	0.692	0.729	0.477	0.699						
1	7,034	13,261	0.726	0.737	0.653	0.728	7,857	18,098	0.650	0.653	0.582	0.639
2	13,261	13,261	0.734	0.738	0.664	0.732	18,098	18,098	0.655	0.656	0.588	0.651
3	19,485	13,261	0.737	0.739	0.671	0.731	28,339	18,098	0.654	0.657	0.595	0.652
4	25,712	13,261	0.738	0.739	0.676	0.734	38,580	18,098	0.653	0.657	0.601	0.653
5	31,939	13,261	0.738	0.739	0.675	0.734	48,821	18,098	0.651	0.658	0.605	0.654
6	38,166	13,261	0.737	0.739	0.681	0.734	59,062	18,098	0.649	0.658	0.606	0.655
7	44,393	13,261	0.737	0.739	0.679	0.735	69,303	18,098	0.647	0.658	0.597	0.656
8	50,620	13,261	0.736	0.739	0.676	0.736	79,544	18,098	0.646	0.658	0.591	0.656
9	56,847	13,261	0.735	0.739	0.674	0.736	89,785	18,098	0.645	0.658	0.588	0.657
10	63,074	13,261	0.735	0.739	0.665	0.736	100,026	18,098	0.644	0.658	0.585	0.657
11	69,301	13,261	0.734	0.739	0.657	0.736	110,267	18,098	0.643	0.658	0.552	0.657
12	75,528	13,261	0.733	0.739	0.645	0.736	120,508	18,098	0.642	0.658	0.518	0.657
13	75,528	19,488	0.736	0.739	0.675	0.737	120,508	28,339	0.646	0.658	0.592	0.657
14	75,528	25,715	0.737	0.739	0.683	0.737	120,508	38,580	0.650	0.658	0.606	0.657
15	75,528	31,942	0.738	0.739	0.681	0.737	120,508	48,821	0.652	0.658	0.612	0.656
16	75,528	38,169	0.738	0.739	0.685	0.739	120,508	59,062	0.653	0.658	0.615	0.655
17	75,528	44,396	0.738	0.739	0.686	0.738	120,508	69,303	0.654	0.658	0.616	0.655
18	75,528	50,623	0.737	0.739	0.683	0.737	120,508	79,544	0.655	0.658	0.615	0.654
19	75,528	56,850	0.737	0.739	0.680	0.736	120,508	89,785	0.656	0.658	0.614	0.655
20	75,528	63,077	0.736	0.739	0.681	0.738	120,508	100,026	0.656	0.658	0.613	0.655
21	75,528	69,304	0.735	0.739	0.677	0.736	120,508	110,267	0.656	0.658	0.614	0.655
22	75,528	75,528	0.735	0.739	0.675	0.737	120,508	120,508	0.657	0.657	0.611	0.656
23	75,528	81,755	0.734	0.739	0.674	0.737	120,508	130,749	0.657	0.657	0.611	0.655
24	75,528	87,982	0.733	0.739	0.674	0.737	120,508	140,990	0.657	0.657	0.611	0.654
25	75,528	94,209	0.733	0.739	0.672	0.737	120,508	151,231	0.656	0.657	0.609	0.654
26	75,528	100,436	0.732	0.739	0.671	0.737	120,508	161,472	0.656	0.657	0.611	0.654
27	75,528	106,663	0.731	0.739	0.671	0.737	120,508	171,713	0.656	0.657	0.609	0.654
28	75,528	112,890	0.730	0.739	0.670	0.737	120,508	181,954	0.656	0.657	0.609	0.654
29	75,528	119,117	0.730	0.739	0.669	0.737	120,508	192,195	0.656	0.657	0.610	0.654

B_2 = standard undersampling (goods = number of bads), B_{12} = the original unbalanced data set, and B_{22} = standard oversampling (bads = number of goods).

For dataset *B*, the optimal balancing occurs at B_{14}, B_{23}, B_{17} and B_{13} for LR, LDA, CART and NN respectively. We suspect that the application of a single over- or under-sampling strategy will be sub-optimal for some sub-regions within the problem domain. For example, it is quite possible that bads are actually the majority class in some regions, and therefore, a more appropriate strategy for this region would be to oversample goods, not bads. This leads us to propose that one further area of study be the application of a regional sub-division algorithm, such as clustering, followed by the application of separate balancings to each of the resulting clusters. Alternatively, a preliminary model could be constructed, with balancing applied based on the posterior probability estimates from the preliminary model.

The results confirm the results of previous studies on related datasets, e.g. on large datasets with strong imbalances in direct marketing (Crone et al., 2006), supporting their validity. In analysing our results, it is apparent that changes to the sample distribution lead to different locations of the decision boundary and classifications of unseen instances, caused by altered cumulative error signals during parameterisation (for a visualisation of shifted decision boundaries, albeit on another aspect of sample selection bias, see e.g. Wu & Hand, 2007). Further evidence of this can be found in the changed coefficients of LR and NN (which may be interpreted directly for a given variable), at times even changing the sign of the coefficient, which one may be less concerned with if one is interested primarily in increases in predictive accuracy. However, this may have implications for the interpretation of the model, and would require a thorough evaluation in practice. Also, possible interactions with initiatives to adjust for reject inference should be evaluated carefully, in order to assess whether they are fully compatible.

4.3. Joint effect of sample size and balancing

Stage 3 considered the joint effect of varying the sample size and balancing in combination. Fig. 3 shows the relative performances of LR, LDA, CART and NN for undersampling (B_2), the unbalanced data set (B_{12}) and oversampling (B_{13}) for increasing sample sizes.

Fig. 3 displays a number of features. The first is that the sample size clearly has an effect on the relative performances of different balancings. In particular, for smaller sample sizes, undersampling performs poorly across both data sets for LR, LDA and NN. The relative performance of undersampling compared to oversampling shows a monotonic increasing trend as the sample size increases, until the difference in performance for LR, LDA and NN becomes small for the largest samples. However, at no point does undersampling ever outperform oversampling for these three methods. In addition, for NN and LR, undersampling marginally underperforms the unbalanced data set for all sample sizes. For LDA, the

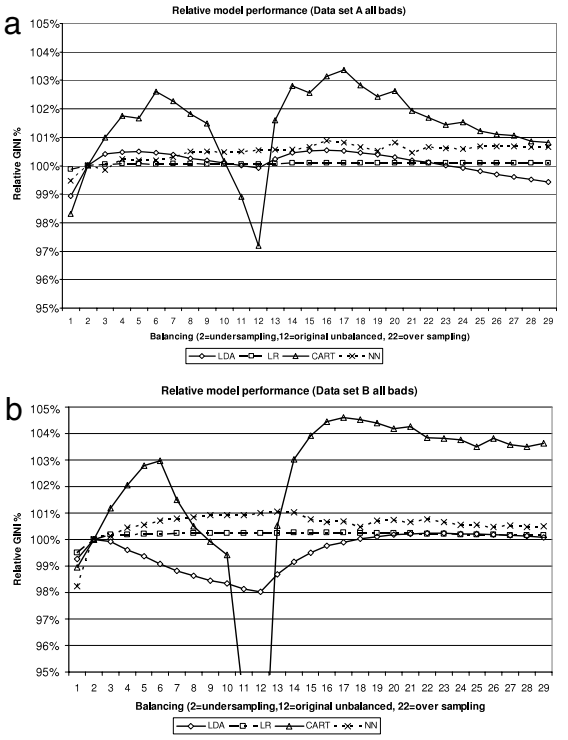

Fig. 2. Absolute model performances for datasets *A* (a) and *B* (b) using all available data ($p = 100$).

story is somewhat different. In general, oversampling outperforms the unbalanced data set, but for smaller sample sizes, undersampling performs poorly once again. CART shows the most divergent behaviour between methods. In particular, the best method for balancing the data appears to be very dependent upon the sample size. For small sample sizes, undersampling performs well, but the performance of oversampling relative to that of undersampling increases monotonically as the sample size increases, until a point is reached at which the situation is reversed, with oversampling being superior for larger samples.

5. Conclusions and discussion

This paper has addressed two issues, sample size and balancing. For the sample size, the position adopted in practice by many scorecard developers is that a sample containing around 1500–2000 cases of each class (including any holdout sample) is sufficient for building and validating a credit scoring model which is nearly optimal in terms of predictive performance. The results presented in this paper undermine this view, having demonstrated that there are significant benefits from taking samples many times larger than this. As a consequence, the paper challenges current beliefs by suggesting the use of significantly larger samples then those commonly used in credit scoring practice and academic studies, even for the well researched LR, contributing to the current discussion on data preprocessing and modelling for credit scoring.

A further issue in relation to the sample size relates to the relative efficiency of algorithms. The results presented in this paper support the case that efficient modelling techniques, such logistic regression, obtain near a optimal performance using far fewer observations than methods such as CART and NN. Therefore, the sample size should be considered when deciding which modelling technique to apply.

Another practice which is widely adopted by scorecard developers is undersampling. Equal numbers of goods and bads are used (by excluding instances of the majority class of goods) for model development, with weighting being applied so that the performance metrics are representative of the true population. Our experiments provide evidence that oversampling significantly increases the accuracy relative to undersampling, across all algorithms, a novel insight which confirms prior research in data mining for imbalanced credit scoring datasets (albeit at the cost of larger datasets and longer training times, and hence reduced resource efficiency). For logistic regression, the most popular technique used for constructing credit scoring models in practice, the balancing applied to datasets appears to be of minor importance (at least for modestly imbalanced data sets such as the ones discussed in this paper). However, the other methods demonstrate a greater sensitivity to balancing, particularly LDA and CART, where oversampling should be considered as the new best practice in assessing them as contender models to LR.

The results hold across two datasets in credit and behavioural scoring, indicating some level of consistency of the results. Here, the choice of two heterogeneous datasets reflects an attempt to assess the validity of the findings across different data conditions, rather than an attempt to increase the reliability. However, while one should be careful to generalise experimental findings beyond the properties of an empirical dataset, credit scoring datasets are remarkably similar across lenders and geography, and might yield more representative results if controlling for the sample size and balance. However, in the absence of additional datasets of sufficient size, the obvious limitations of any empirical ex-post experiment remain.

With regard to further research, there are a number of avenues for further study. One area is the application of active learning (Cohn, Atlas, & Ladner, 1994; Hasenjager & Ritter, 1998), by selecting cases of imbalanced classes that provide a better representation of both sides of the problem domain during the parameterisation phase, promising smaller samples with similar levels of performance to those of larger random samples. Also, there is evidence that instance sampling may have interactions with other preprocessing choices which occur prior to modelling (Crone et al., 2006). Consequently, popular techniques in credit scoring which employ, for example, weights of evidence (Hand & Henley, 1997; Thomas, 2000), instead of the pure dummy variable categorization evaluated here, must be evaluated on different forms of over- and undersampling.

The conclusions drawn from the experiments in instance sampling have implications for previous research findings. In general, previous studies in credit scoring have not reflected the recommendations employed in practice,

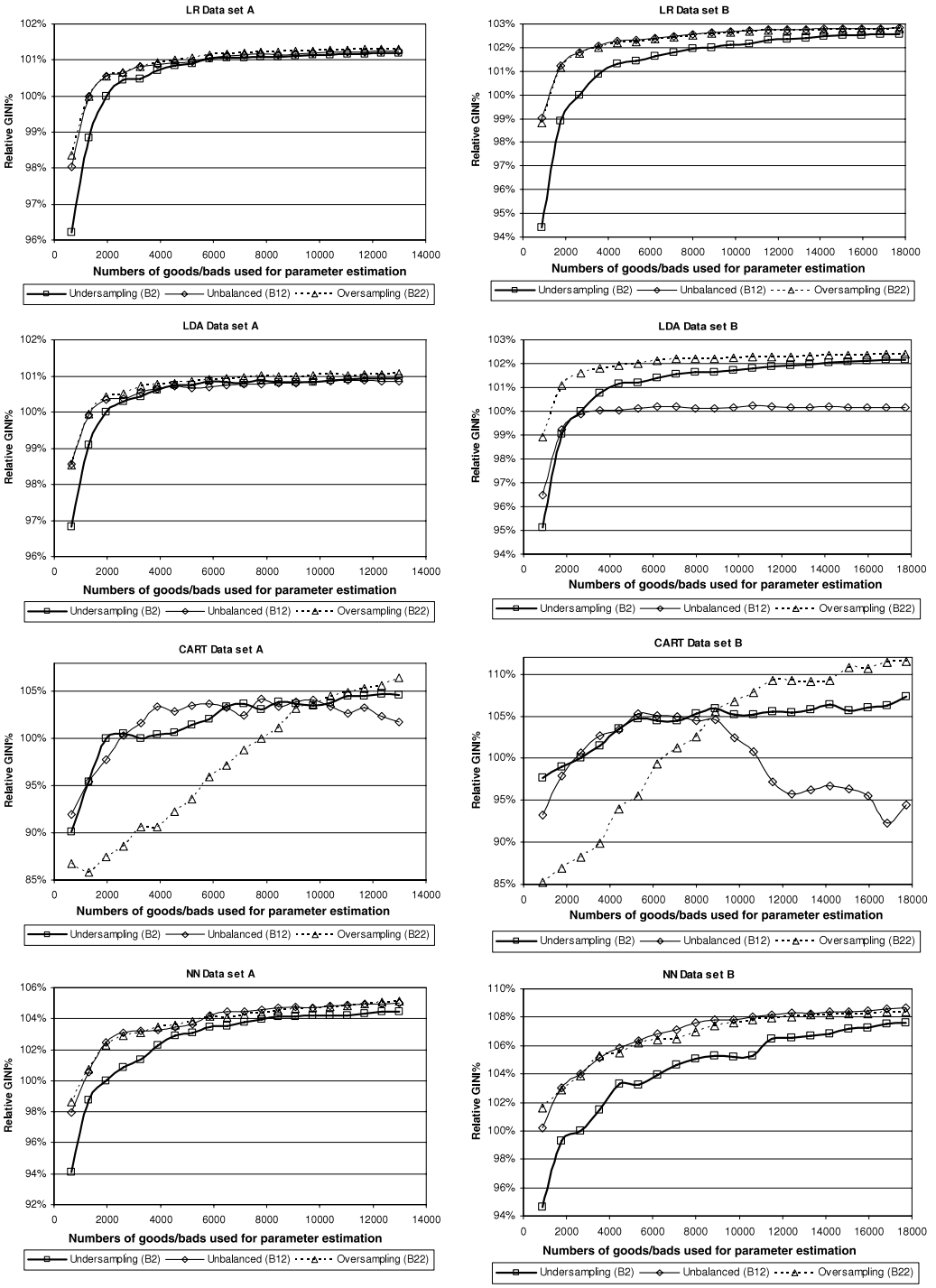

Fig. 3. Effect of balancing in combination with sample size.

evaluating small and imbalanced datasets, which calls into question the validity and reliability of their findings on real-world datasets. Replication studies which re-evaluate these empirical findings across different sampling strategies may resolve this discrepancy. Similarly, the relatively few academic studies of sub-population modelling applied to credit scoring have come to somewhat mixed conclusions, yet the practice is widely accepted in industry, where it is more common to employ larger samples. Considering the smaller sample sizes employed across most academic studies, our research would identify this as a limiting factor which could also explain why sub-population models have failed to show better levels of performance than might have been expected (see e.g. Banasik, Crook, & Thomas,

1996). Therefore, in revisiting the data conditions of prior academic studies through replication, enhanced and novel experimental results may be achieved using an increased sample size and different balancing, that could yield further insights into increasing predictive accuracy for credit scoring practices.

References

Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford: Oxford University Press.

Arminger, G., Enache, D., & Bonne, T. (1997). Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis, and feedforward networks. *Computational Statistics*, *12*(2), 293–310.

Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(5), 627–635.

Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, *183*(3), 1582–1594.

Banasik, J., Crook, J. N., & Thomas, L. C. (1996). Does scoring a sub-population make a difference? *International Review of Retail, Distribution and Consumer Research*, *6*(2), 180–195.

Barclay's (2008). Barlcay's annual report 2008. http://www.barclaysannualreport.com/ar2008/files/pdf/Annual_Report_2008.pdf (accessed on 9.12.09).

Boyle, M., Crook, J. N., Hamilton, R., & Thomas, L. C. (1992). Methods applied to slow payers. In L. C. Thomas, J. N. Crook, & D. B. Edelman (Eds.), *Credit scoring and credit control*. Oxford: Clarendon Press.

Chawla, N. V. (2003). C4.5 and imbalanced datasets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *Proceedings of the ICML'03 workshop on class imbalances*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357.

Chawla, N. V., Japkowicz, N., & Kolcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD: Explorations*, *6*(1), guest editors.

Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, *15*(2), 201–221.

Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, *173*(3), 781–800.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*(3), 1447–1465.

Desai, V. S., Conway, D. G., Crook, J., & Overstreet, G. (1997). Credit-scoring models in the credit union environment using neural networks and genetic algorithms. *IMA Journal of Mathematics Applied in Business and Industry*, *8*(4), 323–346.

Drummond, C., & Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Proceedings of the ICML'03 workshop on learning from imbalanced data sets*.

Evans, D., & Schmalensee, R. (2005). *Paying with plastic. The digital revolution in buying and borrowing*. Cambridge, Massachusetts: The MIT Press.

Finlay, S. M. (2006). Predictive models of expenditure and indebtedness for assessing the affordability of new consumer credit applications. *Journal of the Operational Research Society*, *57*(6), 655–669.

Finlay, S. (2008). *The management of consumer credit: theory and practice*. Basingstoke, UK: Palgrave Macmillan.

Fox, J. (2000). *Nonparametric simple regression*. Newbury Park: Sage.

Hand, D. J. (2005). Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society*, *56*(9), 1109–1117.

Hand, D. J. (2009a). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, *77*, 103–123.

Hand, D. J. (2009b). Mining the past to determine the future: problems and possibilities. *International Journal of Forecasting*, *25*, 441–451.

Hand, D. J., & Henley, W. E. (1993). Can reject inference ever work? *IMA Journal of Management Mathematic*, *5*, 45–55.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *160*(3), 523–541.

Harrell, F. E., Jr., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*(4), 361–387.

Hasenjager, M., & Ritter, H. (1998). Active learning with local models. *Neural Processing Letters*, *7*(2), 107–117.

Henley, W. E. (1995). *Statistical aspects of credit scoring*. Milton Keynes: Open University.

Jentzsch, N. (2007). *Financial privacy: an international comparison of credit reporting systems*. New York: Springer.

Kim, Y., & Sohn, S. Y. (2007). Technology scoring model considering rejected applicants and effect of reject inference. *Journal of the Operational Research Society*, *58*(10), 1341–1347.

Laurikkala, J. (2002). Instance-based data reduction for improved identification of difficult small classes. *Intelligent Data Analysis*, *6*(4), 311–322.

Lewis, E. M. (1992). *An introduction to credit scoring*. San Rafael: Athena Press.

Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, *56*, 1099–1108.

Lovie, A. D., & Lovie, P. (1986). The flat maximum effect and linear scoring models for prediction. *Journal of Forecasting*, *5*(3), 159–168.

Maloof, M. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proceedings of the ICML'03 workshop on learning from imbalanced data sets*.

Mays, E. (2001). *Handbook of credit scoring*. Chicago: Glenlake Pub. Co. Fitzroy Dearborn Pub.

McNab, H., & Wynn, A. (2003). *Principles and practice of consumer risk management*. The Chartered Institute of Bankers.

Miller, M. J. (2003). *Credit reporting systems and the international economy*. Cambridge, Massachusetts: The MIT Press.

Ong, C. S., Huang, J. J., & Gwo-Hshiung, T. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications*, *29*(1), 41–47.

Piramuthu, S. (2006). On preprocessing data for financial credit risk evaluation. *Expert Systems with Applications*, *30*, 489–497.

Prati, R. C., Batista, G., & Monard, M. C. (2004). Learning with class skews and small disjuncts. In *Lecture notes in computer science*: *Vol. 3731. Advances in artificial intelligence—SBIA 2004* (pp. 296–306). Springer.

Quinlan, J. R. (1992). *C4.5: programs for machine learning*. San Mateo, CA: Morgan-Kaufman.

Siddiqi, N. (2006). *Credit risk scorecards: developing and implementing intelligent credit scoring*. John Wiley & Sons.

Somol, P., Baesens, B., Pudil, P., & Vanthienen, J. (2005). Filter- versus wrapper-based feature selection for credit scoring. *International Journal of Intelligent Systems*, *20*, 985–999.

Steyerberg, E. W., Eijkemans, M. J. C., Harrell, F. E., Jr., Habbema, J., & Dik, F. (2000). Prognostic modelling with logistic regression analysis: a comparison of selection methods in small data sets. *Statistics in Medicine*, *19*(8), 1059–1079.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining* (1st ed.) Boston: Pearson Addison Wesley.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149–172.

Thomas, L. C., Banasik, J., & Crook, J. N. (2001). Recalibrating scorecards. *Journal of the Operational Research Society*, *52*(9), 981–988.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Philadelphia: SIAM.

Thomas, L. C., Oliver, R. W., & Hand, D. J. (2005). A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society*, *56*(9), 1006–1015.

Verstraeten, G., & Van den Poel, D. (2005). The impact of sample bias on consumer credit scoring performance and profitability. *Journal of the Operational Research Society*, *56*(8), 981–992.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, *6*(1), 7–19.

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, *19*, 315–354.

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*, *27*(11–12), 1131–1152.

Wu, I.-D., & Hand, D. J. (2007). Handling selection bias when choosing actions in retail credit applications. *European Journal of Operational Research*, *183*, 1560–1568.

Sven Crone is an Assistant Professor (Lecturer) in Management Science at the Lancaster University Management School, UK. He is also deputy director of the Lancaster Research Centre for Forecasting. His main research interests are in forecasting and data mining with artificial neural networks and support vector machines.

Steven Finlay has worked in the field of consumer credit for over fifteen years, both as an academic and as a practitioner. His research interests cover all areas of consumer credit, forecasting and data mining. He is currently the head of analytics at HML in the UK and is a visiting research fellow at Lancaster University, also in the UK.