Instance Sampling in Credit Scoring: an empirical study of sample size and balancing

Journal:	*International Journal of Forecasting*
Manuscript ID:	INTFOR_0912294.R2
Manuscript Type:	Original Article
Keyword:	Credit scoring, Data pre-processing, Sampling, Undersampling, Oversampling, Balancing, Consumer credit
Abstract:	To date, best practice in sampling credit applicants has been established based largely on expert opinion, who generally recommend that small samples of 1,500 instances of both goods and bads are sufficient, and that the heavily biased datasets should be balanced by undersampling the majority class. Consequrntly, the topic of sample size and sample balance has not been subject to formal study in credit scoring, nor to empirical evaluation across different data conditions and algorithms of varying efficiency. This paper describes an empirical study on instance sampling in predicting consumer repayment behaviour, evaluating the relative accuracy 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 oversampling the goods and bads respectively. The paper makes a practical contribution to model building on credit scoring datasets, and provides evidence that using larger samples than those recommended in credit scoring practice provides a significant increase in accuracy across algorithms.

Instance Sampling in Credit Scoring:

an empirical study of sample size and balancing

Abstract: To date, best practice in sampling credit applicants has been established based largely on expert opinion. Helped by the similar properties of data sets across different lenders and geo-graphical regions, they generally recommend that small samples of 1,500 instances of both goods and bads are sufficient, and that the heavily biased datasets should be balanced by undersampling the majority class. In contrast, research on sampling imbalanced datasets in data mining suggests that undersampling yields inferior classification accuracy (for some algorithms), so oversampling the minority class should be conducted. Furthermore, academic studies have employed even smaller datasets, without balanced sampling to retain the original class imbalance. Despite these apparent discrepancies, the topic of sample size and sample balance has not been subject to formal study in credit scoring, nor to empirical evaluation across different data conditions and algorithms of varying efficiency. This paper describes an empirical study on instance sampling in predicting consumer repayment behaviour, evaluating the relative accuracy 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 oversampling the goods and bads respectively. Beyond the common agreement that a larger sample size is beneficial to model construction, and that smaller sample sizes are more resource efficient, the paper makes a practical contribution to model building on credit scoring datasets, and draws several novel conclusions: (a) Using larger samples than those recommended in credit scoring practice and academic studies provides a significant increase in accuracy across algorithms, so that recommendations as to an efficient increase in size can be made. (b) Balanced data sets created via random oversampling out-perform those created

using random undersampling and datasets showing the original class imbalances, contrary to recommendations in practice and those used in most research studies respectively. (c) Different algorithms show different sensitivity to sample size and the choice of balancing, providing an explanation of the inconsistency of their relative performance rankings across past studies caused by the arbitrary sample sizes and sample imbalances of the evaluated datasets.

Keywords: Credit scoring, Data pre-processing, Sample size, under-sampling, over-sampling, Balancing.

1. *Introduction*

The vast majority of consumer lending decisions, who to grant credit to or not, are made using automated credit scoring systems based on an individual's credit score. Credit scores provide an estimate of an individual being a "good" or "bad" credit risk (i.e. a binary classification), which are generated using predictive models of repayment behaviour of previous credit applicants whose repayment performance has been observed for a period of time (Thomas et al. 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 customer databases often proved ineffective or inefficient (given cost, resource and time constraints) to develop predictive models using the complete customer database. As a consequence, 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 to construct the samples which are later used to build accurate and reliable credit scoring models.

Despite its apparent relevance, 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 on their considerable experience of scorecard development. With regard to a suitable sampling strategy, both propose random undersampling to address class imbalances, and suggest that a sample containing

1,500-2,000 instances of each class (including any validation sample) should be sufficient to build 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 model building, these recommendations have not been substantially challenged, neither in practice nor in academic research, which has rather focussed on comparing accuracy of different predictive algorithms on even smaller and unbalanced datasets. As a consequence, issues of sample size and balancing have been neglected as a topic of study within the credit scoring community.

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

consuming aspect of any scorecard development (Anderson, 2007), this omission is surprising and indicates a significant gap in research.

In contrast to credit scoring, issues of sample imbalances have received substantial attention in data mining, leading to the development of frameworks on modelling rare data (Weiss, 2004) and best practices in oversampling through instance resampling and instance creation to balance sampling (Chawla, 2004), which have not been explored in credit scoring. As proven alternatives to instance sampling exist, they warrant a discussion and empirical assessment for credit scoring.

In this paper two aspects of sampling strategy are explored regarding their empirical impact on model performance for datasets of credit scoring structure: sample size and sample balance. Section two reviews prior research, both in best practices and empirical studies, and identifies a gap in research on instance sampling. Both sample size and balance are discussed, reflecting as to whether sample size remains an issue for scorecard developers, given the computational resources available today, and how random oversampling and undersampling may aid in predictive modelling. An empirical study is then described in section three, examining the relationship between sample strategy and 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, and of 29 samples of class imbalances by gradually over- and undersampling the number of goods and bads on each subset respectively. Having looked at both sample size and balancing in isolation, the final part of the paper considers the interaction of sample size and balancing and how predictive performance co-varies with each of

these dimensions. All results of sampling strategy are assessed across four competing classification techniques, which are well established and are known to have practical application 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 the light of differences in the statistical efficiency of estimators with regard to 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) across different classifiers, which may yield explanations for their relative performance, beyond practical recommendations to increasing sample size and / or balancing 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, achieved through data preprocessing (DPP) in the form of data reduction, aiming at decreasing the size of the datasets by means of instance selection and / or feature selection, and data projection, altering the representation of data, e.g. by categorisation of continuous variables. 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.

Following the original advice by Lewis (1992) and Siddiqi (2006), recommendations on sample size concur that 1,500 instances of each class (goods, bads and indeterminates) should be sufficient to build robust models of high quality (see e.g. Anderson, 2007; Mays, 2004; McNab and Wynn, 2003, amongst others). These include data for validation, although fewer cases are required, perhaps a minimum of 300 of each (Mays, 2004). Anderson (2007) justifies their use empirically, as both experts have worked in practice for many years and recommendations appear to be sufficiently large to reduce the effects of multicolinearity and overfitting when working with correlated variables. However, he remarks that no logic was provided for the choice of these numbers, which were determined in the 1960s when the task of collecting data was more costly, and that they have lived on since then without evaluation or challenge (Anderson, 2007), although in practice larger samples are sometimes taken where available (Hand and Henley 1997; Thomas et al. 2002).

The 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 (Thomas et al. 2002; Finlay 2006), and use standardized industry data sources such as those supplied by credit reference agencies. Although credit reference data varies 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 (Miller 2003; Jentzsch 2007). Consequently, datasets are homogeneous regarding the features, i.e. these customer characteristics which hold predictive power. (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 various organisations, from banks and building societies to retailers, other dataset attributes will differ (Hand and Henley, 1997). Dataset sizes, although generally considered 'large', will vary from ubiquitous data on retail credit to fewer customers in wholesale credit (Anderson, 2007). Similarly, sample distributions of datasets, although generally biased towards the goods with relatively fewer bads, will vary to reflect the different risks of the lending decision. Empirical class imbalances range from around 2:1 of goods versus the bads for some sub-prime portfolio's to over 100:1 for high quality mortgage portfolios. It is unclear how recommendations hold across heterogeneous variations of dataset properties.

Table 1 summarizes the algorithms and data conditions of sample size and sample balance from a structured literature review of academic publications that 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. Methods and samples used in empirical studies on credit scoring.

Study	Methods						Dataset & Samples				
	LDA	LR	NN	KNN	CART	Other	data sets	good cases[2]	bad cases[2,3]	goods : bads	indep. vars.
Boyle et al. (1992)	X				X	hyb.LDA	1	662	139	4.8 : 1	7 to 24
Henley (1995)	X	X		X	X	PP, PR	1	6,851	8,203	0.8 : 1	16
Desai et al. (1997)[4]	X	X	X			GA	1	714	293	2.4 : 1	18
Arminger et al. (1997)	X	X	X				1	1,390	1,294	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		1,056	264	4.0 : 1	19
						LP		2,376	264	9.0 : 1	19
								1,388	694	2.0 : 1	33
								3,555	1,438	2.5 : 1	33
								4,680	1,560	3.0 : 1	16
								6,240	1,560	4.0 : 1	16
Ong et al. (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 Disciminant Analysis, LP=Linear Programming, LR =Logistic Regression, NN= Neural Networks, QDA=Quadratic Disciminant Analysis, PP=Projection Pursuit, PR=Poisson Regression, RS=Rough sets, SVM=Support Vector Machines.

[2] In some studies the number of goods/bads used to estimate 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 had occurred.

[4] Three data sets from different credit unions were used. Models were estimated using individual data sets and a combined data set. Figures quoted are for the combined data set.

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) studies in credit scoring have ignored possible DPP parameters of sample size and sample distribution. If sample size and / or sample imbalances were to have a significant impact on predictive accuracy of some algorithms, results across different studies in credit scoring might be impaired. Furthermore, (b) datasets used in academic studies do not reflect the recommendations in practice or practitioner literature, assessing relative accuracy of algorithms across much smaller and imbalanced datasets (of the original sample distribution), questioning the representativeness of prior academic findings for

practice. This echoes similar important omissions in other areas of corporate data mining, such as direct marketing (Crone et al., 2008), and warrants a systematic empirical evaluation across data conditions.

As a further observation, most studies seem preoccupied with predictive accuracy, but fail to reflect other objectives such as interpretability and resource efficiency (in time and costs), which also determine the empirical adequacy of different algorithms in practice. Beyond accuracy, the interpretability of models - and subsequently whether the model is in line with the intuition of the staff - is often of even greater importance; speed (of classification itself and with which a score-card can be revised) and robustness are also of relevance (see e.g. Hand and Henley, 1997). Methods of computational intelligence, 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 widely used 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 (Thomas et al. 2005; Crook et al. 2007), 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, limited operational requirements and particularly their interpretability (Finlay 2008). As DPP choices in sample size and balance may not only impact 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 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 and Henley (1993)).

2.2. Sample size

Instance sampling is a common approach in statistics and data mining, 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 et al., 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 discriminatory power of a model built on a sample is comparable to 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 will also eliminate much of the advantages that sampling provides, such as reduced computation time and costs of data acquisition. In smaller samples, patterns contained in the data may be missed or erroneous patterns be detected, enhancing efficiency at the costs of limiting accuracy. Therefore, determining an efficient and effective sample size requires a methodical approach given the properties of the datasets, 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 differ 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 and Schmalensee 2005). Samples are conventionally built using a stratified random sample (without replacement) on the target variable, drawing an equal number of goods and bads proportional 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 size is often the number of bads, with few organizations having more than a few thousand, or at most tens of thousands, of bad cases (Andersen, 2007).

From a theoretical perspective, the issue of using larger sample sizes might at first sight appear to have been resolved in recent years due to the increased computational resources provided by a modern PC. Credit scoring models can be developed using popular approaches such as LR, LDA and CART using large samples of observations within a few hours. As using the complete dataset available, rather than further sampling from it, has become a feasible course of action for most consumer credit portfolios, it is valid to ask: does sample size remains a relevant issue for scorecard development? In practical situations, sample size remains an important issue for multiple reasons. First, there are often costs associated with acquiring data from a credit reference agency, resulting in a trade-off between an increase in accuracy obtained from using larger samples and the marginal cost of additional data acquisition. Also, a model developer may rebuild a model many, possibly dozens of times, to ensure that the model meets business requirements that are often imposed on such models, or to evaluate the effect of different DPPs or different meta-parameters of algorithms 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 a large and significant reduction in 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 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 and Henley, 1997), not all records of past applicants are representative of current / future behaviour and hence require sampling. This is particularly apparent in the recent credit crunch, where new models needed to be constructed for novel economic circumstances (Hand, 2009), limiting the ability to use all data and asking the question of a minimal (or near optimal) sample size for a renewed scorecard development. 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 of sample size.

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

2.3. Sample distribution (Balancing)

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

The importance of reflecting the imbalance between the majority and minority class 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 cost of a type I vs. type II error is dramatically asymmetrical, making an invalid prediction of the minority class more costly than accurate prediction of the majority class. Yet traditional classification algorithms - driven by the objective to minimise some loss of error function across two different parts of a population - typically have a bias towards the majority class that provides more error signals. Therefore, the underlying problem requires 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 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 and Provost, 2003). In undersampling, instances of the minority and majority class are randomly selected 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 underrepresented class are replicated a number of times, so that the class distributions are more equal. Note that inconsistencies in the 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 leads to models with enhanced discriminatory power but both random oversampling and random undersampling methods have shortcomings: random undersampling can discard potentially important cases from the majority class of the sample (the goods), impairing an algorithm's ability to learn the decision boundary; random oversampling duplicates identical records and can lead to overfitting of similar instances. (Note that under and oversampling are only conducted on the training data used for model development – 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 observations belonging to the minority class (Weiss 2004). As both over- and undersampling can potentially reduce accuracy in generalising for unseen data, a number of studies have compared variants of over and undersampling, and have presented (often conflicting) viewpoints on the gains in accuracy derived oversampling versus undersampling (Prati et al. 2004; Chawla 2003; Drummond and Holte, 2003; Maloof 2003), indicating that the results are not universal and depend on the dataset properties and the application domain.

However, as datasets in credit scoring share similar properties across lenders, findings on over- vs. undersampling are expected to be more representative across databases.

Reflecting on best practices and empirical studies (see 1.1.), credit scoring practices actively recommends 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 approaches on oversampling 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 oversampling have been developed, e.g. to selectively undersample unimportant instances (Laurikkala, 2001) or to create synthetic examples in oversampling (Chawla, 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- vs. undersampling. It should be noted that both under- and oversampling are later assessed 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, but are suspected to be complimentary to the evaluated choices of instance sampling, which merely provide homogeneous error signals for the information contained within the original sample, and not to augment for missing parts of the population.

Moreover, the error signals derived from different numbers of goods and bads may shift the decision surface in feature space for those methods estimating decisions 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 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 then derive $P(g|\bullet)$ (see Hand and Henley (1993) 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), using two independent output nodes to assess $p(\bullet|g)$ and $p(\bullet|b)$ to derive $P(g|\bullet)$ as in LDA, or combinations by linking multiple output nodes using the softmax function, which pose undefined statistical properties and efficiency. Should these meta-parameter choices impact estimator efficiency, together with increasing number of parameters to be estimated through latent variables, different practical recommendations for an effective dataset size and balance may be the result. As a consequence, asymptotically, we expect an efficient estimator such as LR to remain largely robust to some sample variations, leaving the accuracy of balanced or imbalanced datasets rather unaffected by over- or undersampling, while the parameters of algorithms such as LDA might be fundamentally altered in magnitude (and sometimes in sign), resulting in differences in predictive accuracy and relative ranking to other inefficient algorithms. (It should be noted, that altered parameters may also impact the interpretability of algorithm parameters, which holds potential implications for constructing credit scoring models in practice, these require attention beyond impacts on predictive accuracy.)

The effect of balancing on estimators of different statistical efficiency should be assessed separately to the effect of sample size, and the joint effect of sample distribution and sample size. This assessment will be problematic for strong undersampling, as in creating very small samples the parameters (even of LR) still remain unbiased but may deviate somewhat from the population value due to the inherent variance in smaller samples, despite ensuring random sampling all throughout the experiments. To reflect this, our experiments will include also small sample sizes, but 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 caused predominantly by experimental biases introduced through the arbitrarily chosen small sample size, but not the classifiers' capabilities (i.e., non-linearity etc.) or the sample distribution of the data. Such findings would confirm previous studies, and hence add to the reliability of our findings, and place our assessment of sample size and balance in context to existing research.

Furthermore, it should be noted that over- and undersampling will not only impact predictive accuracy depending on the statistical efficiency, but also resource efficiency in model construction and application. Balancing impacts 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 for model parameterisation (although this seems less important in contrast to improving accuracy, as the time for applying an estimated model will remain unchanged).

3. Experimental Design

3.1. Datasets

Two data sets, both 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 data set (data set 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 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 any time within the last 6 months were classified as bad. This is consistent with 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 (Lewis 1992; Hand and Henley 1997; McNab and Wynn 2003). After removal of outliers and indeterminates the sample contained 88,789 observations of which 75,528 where 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 number, value and time since most recent CCJ/bankruptcy, current and historical account performance, recent credit searches, Electoral Roll and MOSAIC postcode level classifiers.

The second data set, data set B, was a behavioural scoring data set from a mail order catalogue retailer providing revolving credit. Performance data was 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 no more than one month in arrears, bads as 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 bad cases. Data Set 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 sample size on 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 follow procedure:

> Step 1. The population of N observations, comprising G goods and B bads (N=G+B) was segmented into k folds of equal size and p percentiles within each fold. Stratified random sampling was applied, with goods and bads sampled independently to ensure class priors in each fold 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} and N_{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 fold were used to evaluate 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 sample size variation on accuracy, leading to 20 different sample sizes. For a relationship between sample size and accuracy we would expect statistically significant, but also consistent results of increasing accuracy (i.e., a monotonically increasing trend in improving performance for the results to be considered reliable) beyond the recommended "best practice" sample size of 1,500 to 2,000 bads. The number of percentiles was chosen under the constraint of available observations and the number of variables, so that all variables would still contain significant numbers of observations and allow stable parameter estimates when samples sizes were small (a minimum of 250 bads and 500 goods for p=5). To comply with what is reported as standard practice within the credit scoring community, balanced data sets were used with goods randomly under-sampled (excluded) from each fold for model development, so that the number of goods and bads were the same.

3.3. Balancing Sample Distributions

The second part of the study considered balancing. In data mining in general, studies have been conducted using either undersampling, the original distribution of the population, or oversampling on algorithms of varying statistical efficiency. However,

this does not allow inference on possible systematic and continuous effects from 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 different balancing applied for 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 under-sampled data set observations where 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 imbalances, with B_3 representing standard undersampling with goods sampled down to equal the number of bads, and B_2 undersampling the goods beyond the number of bads (i.e., less goods than bads). B_{13} - B_{22} were randomly oversampled with increasing class imbalances, with sample B_{22} representing standard over-sampling with bads re-sampled so that the number of goods and bads were equal. For B_{23}-B_{29} oversampling was extended further, so that the samples contained more bads than goods.

The creates a continuous, gradually increasing imbalance from extreme undersampling to extreme oversampling, spanning most sampling balances employed in data mining while allowing 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 INT(N_{pv}/N_{pb}) times, where N_{pv} is the desired number of bads in the sample (So for standard over-sampling, where the number of bads is equal to the number of goods, $N_{pv} = N_{pg}$). Additional (N_{pv} - 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 same *k*-fold development/validation methodology as described in section 3.1 was adopted, with observations assigned to the same 50 folds. Note that for all experiments no balancing was applied to the test fold; i.e. the class priors within the test fold 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 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 allowed a 3-D surface to be plotted showing how sample size, balancing and performance co-vary, and makes it possible to consider trade-offs between sample size and balancing. It is noted that Part 3 represents a superset of experiments, containing all the those described in parts 1 and 2, as well as many others. We've taken this approach, building up the results in stages, to aid the readability of the paper.

3.4. Methods, Data Pre-Processing and Variable Selection

Methods were chosen to represent those established in credit scoring, including LR, LDA and CART, and NN, a frequently evaluated contender that has shown enhanced accuracy in fraud detection and other back end decision processes (where limited explicability is required, see e.g. Hand, 2001) but so far failed to prove its worth in

credit scoring. As the evaluation of different modelling techniques is not of primary interest in this study, recently developed methods such as SVM etc. are not assessed.

Experiments were repeated for LR, LDA, CART and NN. For 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, and then pruning applied using the 20 percent validation sample as advocated by Quinlan (1992). Binary splits were employed, based on maximum entropy. For NN a MLP architecture was adopted with a single hidden layer. Preliminary experiments were performed to determine the number of hidden units for the hidden layer using the smallest available sample size (i.e. p=5) to ensure that over fitting did not result for small sample sizes. T-1 exploratory models were created using 2,3,…,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 model performance on the 20 percent test sample. Given the size and dimensionality of the data sets involved and the number of experiments performed, we employed a quasi-newton algorithm with a maximum of 100 training iterations to allow experiments to be completed in realistic time.

The most widely adopted approach to pre-processing credit scoring data sets is to categorize the data using dummy variables (Hand and Henley 1997), which generally provides a good linear approximation to 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 et al. (2005) suggest that in most cases between 3 and 6 dummies should be sufficient, although a greater number of dummies may be defined if sufficient volume of data is available.

For data sets A and B, the independent variables were a mixture of categorical, continuous and semi-continuous, which were coded as dummy variables for LDA, LR, NN. All dummy variables for both data sets contained in excess of 500 good and 250 bad cases, and more than 1,000 observations in total (for p=100). For CART preliminary experiments showed the performance based on dummy variables was extremely poor, and better performance resulted from creating an ordinal range using the dummy variable definitions. Therefore, for CART this ordinal categorization was used. We note that this particular data preprocessing may have biased results against some of the nonlinear algorithms {Crone, 2010, ISI:000279134100043}, but it was chosen due to its prevalence in credit scoring practice and academic studies. To allow replication of the experiments, additional details on data pre-processing and method parameterisation are provided from the authors upon request.

3.5. Performance evaluation

To measure model accuracy, a precise estimate of the likelihood of class membership may serve as a valid objective in parameterisation, but it is of secondary importance to a model's ability to accurately discriminate between the two classes of interest (Thomas et al. 2001). As a consequence, measures of group separation, such as the Area under the ROC curve (AUC), the GINI coefficient, KS statistic etc. are widely used to assess 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 class distribution given the data properties in credit scoring (as such, e.g. simple classification rates may not be applied). A popular metric in data mining, the AUC provides a single valued performance measure [0; 1] that assesses the tradeoff between hits vs. false alarms, where random variables score 0.5. To employ a

performance measure more common in the practice of the retail banking sector, we assess model performance using the related GINI coefficient, calculated using the brown formula (Trapezium rule):

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

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, 1997), measuring two times the area between the ROC-curve and the diagonal (with AUC = (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 accuracy of probability estimates for 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 in Hand, 2005), and which adequately reflects our empirical modeling objective. Furthermore, it allows direct comparison of results to other studies, including applications in retail banking where practitioners regularly employ GINI, which is considered equally important. Therefore, despite recent criticism (see e.g. Hand, 2005, 2009), the limited theoretical weaknesses of GINI seem outweighed by its advantages in interpretability by practitioners and across other studies.

4. Experimental Results

4.1. Effect of Sample size

The first stage of 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. Absolute GINI by sample size for data set A and for data set B.

	Dataset A					Dataset B				
p	#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%	1326	0.721 **	0.721 **	0.605 **	0.692 **	1809	0.635 **	0.633 **	0.542 **	0.600 **
15%	1989	0.727 **	0.730 **	0.634 **	0.701 **	2714	0.641 **	0.640 **	0.548 **	0.605 **
20%	2652	0.729 **	0.733 **	0.638 **	0.707 **	3619	0.646 **	0.645 **	0.556 **	0.614 **
25%	3315	0.730 **	0.733 **	0.634 **	0.711 **	4524	0.649 **	0.648 **	0.567 **	0.624 **
30%	3978	0.731 **	0.735 *	0.637 **	0.717 **	5429	0.649 **	0.649 **	0.574 **	0.624 **
35%	4641	0.732 *	0.736 *	0.638 **	0.722 **	6334	0.650 **	0.650 **	0.572 **	0.628 **
40%	5304	0.733	0.736 *	0.644 **	0.723 **	7239	0.651 **	0.651 **	0.572 **	0.633 **
45%	5967	0.733	0.737	0.648 **	0.725 **	8144	0.652 **	0.652 **	0.577 **	0.635 **
50%	6630	0.733	0.737	0.656 **	0.726 **	9049	0.652 **	0.652 **	0.581 **	0.637 **
55%	7293	0.733	0.737	0.658 **	0.727 **	9953	0.652 **	0.653 **	0.576 **	0.636 **
60%	7956	0.733	0.737	0.654 **	0.729 **	10858	0.653 **	0.653 **	0.577 **	0.637 **
65%	8619	0.733	0.737	0.659 **	0.730 **	11763	0.653 **	0.654 **	0.578 **	0.644 **
70%	9282	0.733	0.738	0.658 **	0.730 **	12668	0.654 **	0.655 **	0.578 **	0.644 **
75%	9945	0.733	0.738	0.656 **	0.731 *	13573	0.654 **	0.655 **	0.580 **	0.645 **
80%	10608	0.734	0.738	0.659 **	0.731 *	14478	0.654 *	0.655 **	0.583 **	0.646 **
85%	11271	0.734	0.738	0.663 *	0.731 *	15383	0.655 *	0.656 *	0.579 **	0.648 **
90%	11934	0.734	0.738	0.663	0.732	16288	0.655	0.656	0.581 **	0.649 **
95%	12597	0.734	0.738	0.664	0.732	17193	0.655	0.656	0.582 **	0.650
100%	13261	0.734	0.738	0.664	0.732	18098	0.655	0.656	0.588	0.651

** performance is significantly different from p=100% at 99% level of significance.
* performance is significantly different from p=100% at 95% level of significance.

Table 2 shows both the comparative level of accuracy between methods, as well as changes in the accuracy for increased sample size for each individual method. Also highlighted in Table 2 are the results from paired t-tests to determine 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] $(5 \le x \le 100)$. Observing monotonically increasing significance of 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 efficiency of different classification algorithms as indicated before, with LR and LDA approaching this level at 5,000 bads already, while NN require more than double the muber of instances (but still don't achieve the accuracy of LR).

Table 2 documents two original findings. First, all methods show a monotonic increase in their predictive accuracy (with minor fluctuations in accuracy for CART), which might be considered unsurprising given the common practical understanding that more data is better. However, accuracy increases well beyond the recommended "best practice" sample size of 1,500 to 2,000 instances of each class. For logistic regression and LDA around 5,000 samples of 'bads' are required before performance is statistically indistinguishable from that resulting from a sample size of p=100 for data set A, and around 15,000 cases for data set B. This results in significantly larger (balanced) datasets of 10,000 and 30,000 instances altogether, far exceeding the recommendations of practice and the experimental design of academic studies. Equally, for CART and NN larger samples are required before a level of accuracy is asymptotically approached, but again datasets of larger magnitude yield improved performance. It is important to note that these tests of significance between samples of size p=x and p=100 should only be considered lower bounds on the optimal sample size. This is 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 that the number of observations within each coarse classed interval were chosen so that when samples sizes were small (i.e. p=5), all variables would still contain sufficient numbers 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 increase in the performance of the resulting models.

The effect of increased sample size on algorithm performance are also illustrated in Figure 1, which shows the relative increase in accuracy of each method indexed in relation to the results that are obtained from using industry best practice

recommendations of 1,500 instances obtained via undersampling (= 100%), for both data set A (figure 1a) and data set B (figure 1b).

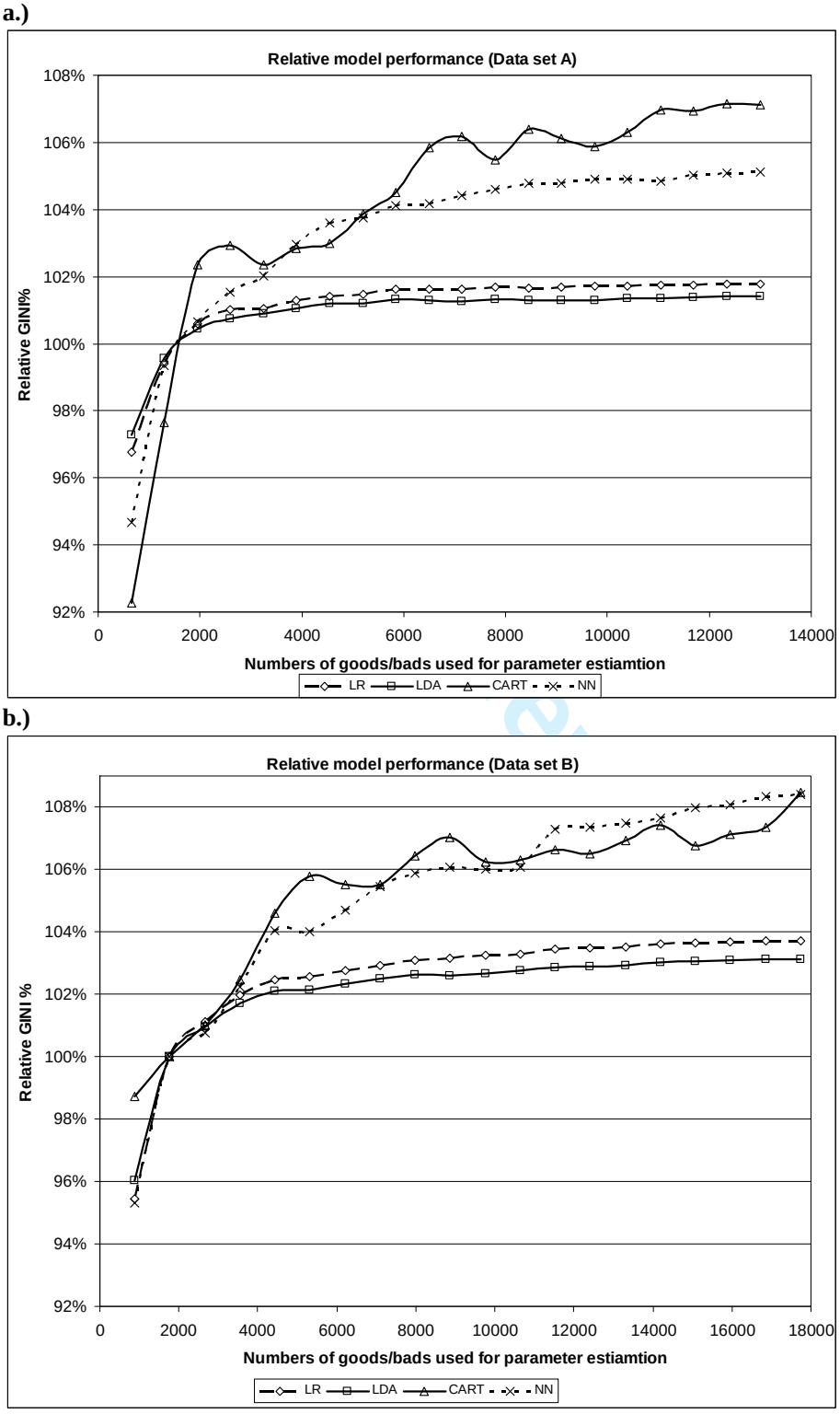


Figure 1. Relative model performance by sample size for data set A (a) and for data set B (b).

Note that Figure 1 provides relative improvement for each method in isolation, and not performance comparisons between the methods. Figure 1 shows similar patterns for models developed using data set A and B, indicating a similar trend in performance with increasing sample size. Increasing sample sizes from 1,500 to the maximum possible, increases 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 improvements are even more substantial, significantly increasing performance by 3.14% for LDA, 3.72% for LR, 8.41% for NN and 8.48% CART (although it should be noted that for both data sets the absolute performance of CART was consistently worse than the other methods – following the trend seen in Table 2). As statistically significant improvements are also feasible from simply increasing sample sizes for the well explored and comparatively efficient estimator of LR, well beyond its best practices, these findings prove novel and significant advancements for credit scoring practice. All increases may be considered substantial considering the flat maximum effect (Lovie and Lovie 1986) and the costs associated with increasing scorecard accuracy by fractions of a percentage point. Furthermore, the differences from sample size are substantial considering the improvements in performance attributed to algorithm choice and tuning in 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 steady as N_{pb} rises above 2,000 and plateaus by around N_{pb} = 5,000 for data set A and 12,000 for data set B, indicating little need for collecting larger datasets. However, for NN and CART it would appear the performance has not plateaued by N_{100b} and therefore, the absolute performance may improve if larger samples were available.

The second prominent feature of note in Table 2 is that the relative performance of different methods varies with sample size, in particular for small samples. One concern that might be raised from the experimental design of 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 et al. 1996), then this suggests that there is a risk of over-fitting where $N_{pb} \leq 810$ (p<=6; i.e. the first row in Table 3 and the first data point in Figure 1). However, the region where $N_{pb} \leq 810$ is not the area of greatest interest. Also, because of the preliminary variable selection procedure, only variables where it is known with a high degree of certainty that a relationship exists between the dependent and independent variables have been included in models. It is also true that for large samples containing hundreds of cases of each class, the ratio of events to variables tends to be a less important factor than for smaller samples (Steyerberg et al. 2000). Therefore, we think it unlikely that over-fitting has occurred.

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

4.2. Effect of Balancing

The second set of analysis reviews the effect on balancing the distribution of the target variable on predictive accuracy. Table 3 provides the results of 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 differences in sample size from rebalancing.

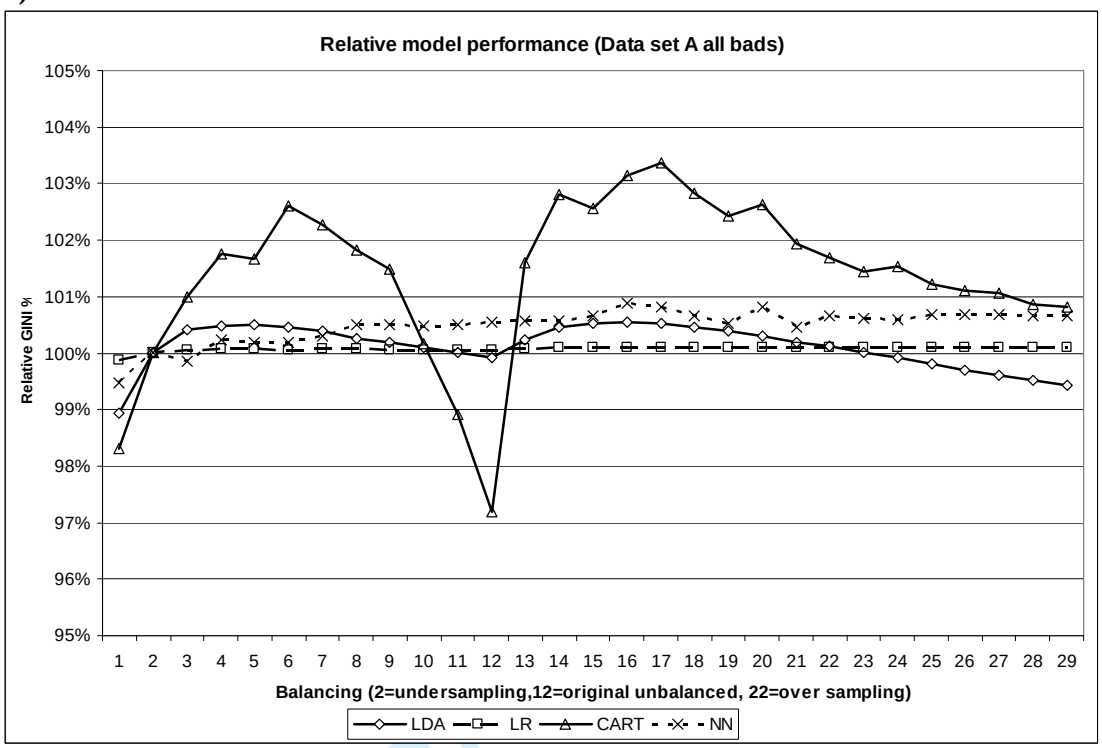
Table 3. Absolute GINI by sample distribution for data set A (a) and for data set B (b).

	Dataset A						**Dataset B**					
B_n	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

B2 = standard undersampling (Goods = number of bads), B12 is the original unbalanced data set, and B22 is standard oversample (Bads = number of goods)

Figure 2 provides a graphical representation of the results.

a.)



b.)

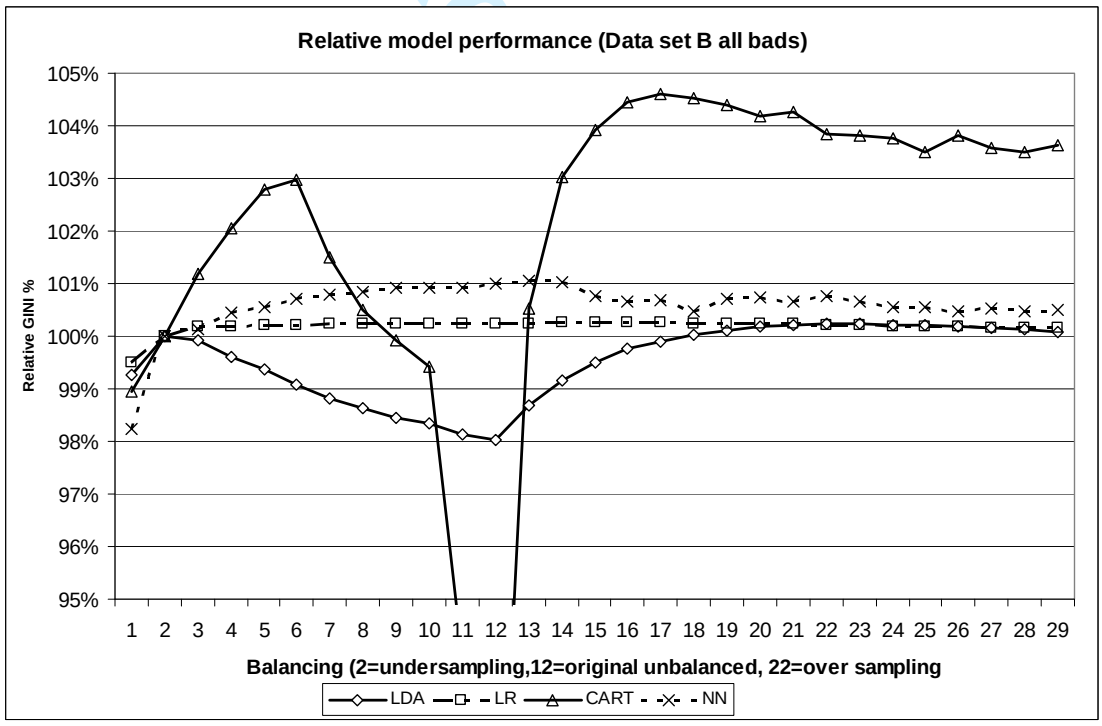


Figure 2. Absolute model performance for data set A (a) and for data set B (b) using all available data (p=100)

In examining the results from Table 3 and Figure 2, we shall begin with the performance of logistic regression. Logistic regression is remarkably robust to

balancing, yielding >99.7% of maximum performance for both data sets, regardless of the balancing strategy applied. For both data sets undersampling leads to worse performance than the unbalanced data set (B_{12}) or oversampling (B_{22}), and using the unbalanced data gives slightly worse performance than oversampling. However, none of these differences are statistically significant. LDA displays greater sensitivity, with performance falling to just under 99.4% maximum for data sets A and 98% for data set B. For both data sets the worst performance for LDA is when B_{12} is applied and these differences are significant at a 99% level of significance. CART is by far the most sensitive technique, with performance of 95% of maximum for data set A and 84% for data set B. NN also shows some sensitivity to balancing, with worse performance, the greater the degree of undersampling applied.

Another, and arguably the most interesting feature displayed in Figures 2, is that maximum performance does not consistently occur at the traditional over-sampling point (B_{22}). For data set A optimal balancing is at B_{21}, $B_{17,}$ B_{18} and B_{16} for LR, LDA, CART and NN respectively. For data set B optimal balancing occurs at B_{14}, $B_{23,}$ B_{17} and B_{13} for LR, LDA, CART and NN respectively. What we suspect is that the application of a single over/under sampling strategy is 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. 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 would 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 posterior probability estimates from the preliminary model.

The results confirm previous studies on related datasets, e.g. on large datasets with strong imbalances in direct marketing (Crone et al, 2005), supporting their validity. In analysing the results, it is apparent that changes to the sample distribution lead to a different location of the decision boundary and classification of unseen instances, caused by altered cumulative error signals during parameterisation (for a visualisation of shifted decision boundaries see e.g. Wu and Hand (2007), albeit on another aspect of sample selection bias). This is further evidenced in changed coefficients of LR and NN (which may directly interpreted for a given variable), at times even changing the sign of the coefficient, which one may be less concerned with if interested primarily in increases in predictive accuracy. However, this may have implications for the interpretation of the model, and would require thorough evaluation in practice. Also, possible interactions with initiatives to adjust for reject inference should be evaluated carefully, to assess if they are fully compatible.

4.3. Joint Effect of Sample Size and Balancing

Stage 3 considered the joint effect of varying sample size and balancing in combination. Figure 3 shows the relative performance of LR, LDA, CART and NN for undersampling (B2), the unbalanced data set (B12) and over-sampling (B13) for increasing sample sizes.

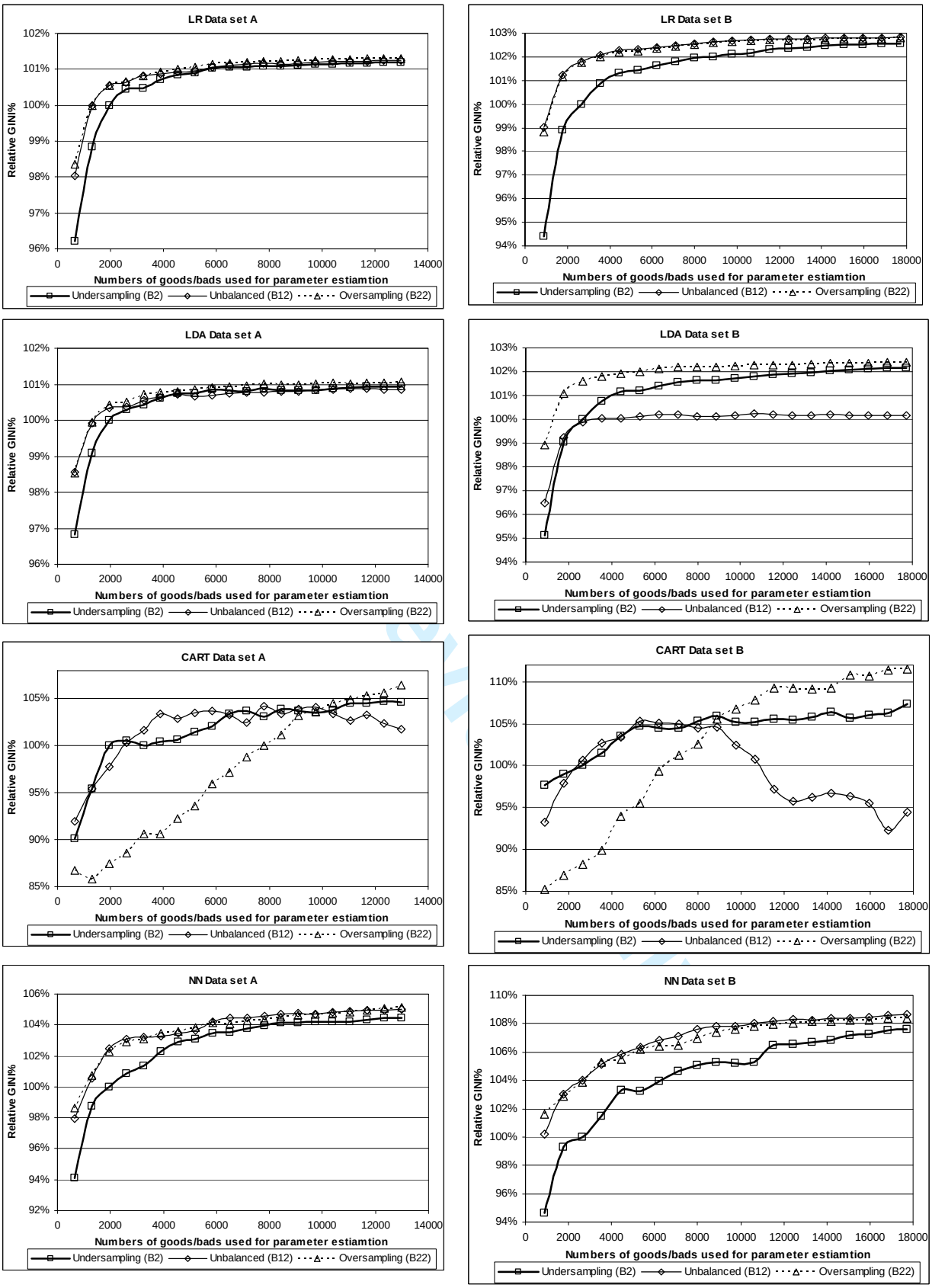


Figure 3. Effect of balancing in combination with sample size

Figure 3 displays a number of features. The first is that sample size clearly has an effect on the relative performance of different balancings. In particular, for smaller sample sizes, undersampling is poor across both data sets for LR, LDA and NN. The relative performance of undersampling compared to oversampling shows a monotonic increasing trend as sample size increases until for the largest samples the difference in performance for LR, LDA and NN becomes small. However, for these three methods, at no point does undersampling ever outperform oversampling. In addition, for NN and LR, undersampling marginally underperforms the unbalanced data set for all sample sizes. For LDA the story is somewhat different. In general, oversampling outperforms the unbalanced data set, but for smaller sample sizes, again undersampling is poor. CART shows the most divergent behaviour between methods. In particular, the best method for balancing the data appears to be very dependent upon sample size. For small sample sizes undersampling performs well, but the relative performance of oversampling compared to undersampling increases monotonically as the sample size increases, until a point is reached at which the situation reverses, with oversampling superior for larger samples.

5. Conclusions and Discussion

This paper has addressed two issues of sample size and balancing. On sample size, the position adopted in practice by many scorecard developers is that a sample containing around 1,500-2,000 cases of each class (including any holdout sample) is sufficient to build and validate a credit scoring model that is near optimal in terms of predictive performance. The results presented in this paper undermine this view, having demonstrated that there are significant benefits to 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 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 optimal performance using far fewer observations than methods such as CART and NN. Therefore, sample size should be a factor that is considered when deciding which modelling technique to apply.

Another practice that is widely adopted by scorecard developers is under-sampling. Equal numbers of goods and bads are used (by excluding instances of the majority class of goods) for model development, with weighting applied so that performance metrics are representative of the true population. Our experiments provide evidence, that oversampling significantly increases accuracy over undersampling, across all algorithms, a novel insight which confirms prior research in data mining also for imbalanced credit scoring datasets. (Albeit, at the cost of larger datasets, longer training times and hence reduced resource efficiency.) For logistic regression, the most popular technique used to construct credit scoring models in practice, the balancing applied to datasets is of minor importance (for modestly imbalanced data sets such as the ones discussed in this paper) due to the procedure's statistical efficiency which results in an insensitivity to the balancing of the training data. However, other methods demonstrate greater sensitivity to balancing, particularly LDA and CART, where oversampling should be considered as a 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 asses the validity of findings across different data conditions, rather than an attempt to increase reliability. However, while one should be careful to generalise experimental findings beyond the properties of an empirical dataset, datasets in credit scoring are remarkably similar across lenders and geography and may yield more representative results if controlling for sample sizes and balances. 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 of further study. One area is the application of active learning (Cohn et al. 1994; Hasenjager and 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 that of larger random samples. Also, there is evidence that instance sampling may have interactions with other pre-processing choices that occur prior to modelling (Crone et al. 2006). Consequently, popular techniques in credit scoring that employ, e.g. weights of evidence (Hand and 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 form the experiments in instance sampling have implications for previous research findings. In general, prior studies in credit scoring did not reflect recommendations employed in practice, evaluating small and imbalanced datasets, which questions the validity and reliability of their findings on real-world datasets. Replication studies that re-evaluate 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 it is a widely accepted practice in industry where larger samples are more commonly employed. Revisiting research on sub-population modelling in the future would suggests that the sample sizes used may have been a key limiting factor in the failure of some sub-population models to show better levels of performance than might have been expected (see e.g. Banasik et al., 1996), and may yield enhanced results using increased sample size and balancing.

Bibliography

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.

Banasik, J., Crook, J.N. & Thomas, L. C. (2001). Scoring by usage. *Journal of the Operational Research Society*, 52(9), 997-1006.

Barclays (2008). Barlcays annual report (2008), http://www.barclaysannualreport.com/ar2008/files/pdf/Annual_Report_2008.pdf. Accessed on 9 December 2009.

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

Chawla, N., Japkowicz, N. & Kolcz, A. (guest editors). (2004). *Special Issue on Learning from Imbalanced Data Sets, ACM SIGKDD Explorations*, 6(1).

Chawla, N. V. 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*, 2003.

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. 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*, 2003.

Evans, D. & Schmalensee, R. (2005). *Paying With Plastic. The Digital Revolution in Buying and Borrowing*, Cambridge Massachusetts: The MIT Press.

Finlay, S. (2008). The *Management of Consumer Credit: Theory and Practice,* Basingstoke, UK: Palgrave Macmillan.

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.

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., & Henley, W. E. (1993). Can reject inference ever work? IMA Journal of Management Mathematics, 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.

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), 159-68.

Maloof. M., 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*, 2003.

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.

Quinlan, J. R. (1992). C4.5: *Programs for Machine Learning*, San Mateo, CA.: Morgan-Kaufman.

Prati, R. C., Batista G. & Monard, M.C. (2004). Learning with class skews and small disjuncts. *Advances in Artificial Intelligence – SBIA 2004*, Lecture notes in computer science Volume 3731. Springer, pp 296-306.

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

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. (2005). Introduction to data mining (1st ed.). Boston: Pearson Addison Wesley.

Tapp, A. (2008). *Principles of Direct and Database Marketing*, FT Prentice Hall.

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.

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