Appears in "Proceedings of Credit Scoring and Credit Control V", Edinburgh, 1997.

Why Size Does Matter in Credit Scoring

Patrick D. Surry[a,b], **Nicholas J. Radcliffe**[a,b]
{pds,njr}@quadstone.co.uk
[a]Quadstone Ltd, 16 Chester Street, Edinburgh, EH3 7RA, UK
[b]Department of Mathematics, University of Edinburgh, King's Buildings, EH9 3JZ, UK

Most scoring models are developed with small historical samples, typically in the region of 5,000–10,000 customers. Modern computing hardware and software allows samples orders of magnitude larger than this to be used. This paper discusses the potential benefits of developing scoring models using the bulk of the customer base, often leading to datasets containing millions of individuals.

Much of the focus of the paper stems from viewing the "regression" or "weight setting" phase of the development of a scoring system as simply one, relatively small, component of a much larger process of building useful models for credit management. This larger process consists of an investigative, exploratory process of repeated refinement in which models are constantly redeveloped in order to ensure that their statistical formulations faithfully reflect the underlying business goals.

Within this view, there is more scope for testing different mathematical definitions of "goods" and "bads", and of the precise predictions to be made by the scoring mechanism—for example, the outcome period over which the likelihood of default is to be modelled, the severity of any predicted losses or the profitability of the account or customer. This requires interactive exploration of the available data, including the ability to test and refine hypotheses and to develop trial models rapidly. From the *business* perspective, it is not only the *predictive* accuracy of the model that is important, but also its *explanatory* power. We will illustrate the way in which interactive visualisation and interrogation can be used to pinpoint and correct weaknesses in scoring models.

It is common practice for the customer base to be segmented and for separate scoring functions to be developed for each segment, but such segmentations are usually rather static, with each model being developed essentially independently. Again, the ability to handle the entire customer base supports a more integrated view of the entire scoring function represented by a segmentation and the models within it and allows this ensemble of models to be developed and managed together. The current drive towards customer-based scoring, as distinct from product- or account-based scoring, also inevitably requires larger datasets to be considered, not least because of the variety of product mixes typically held by different customers.

Even within the limited context of scorecard development, larger samples can be a more powerful tool than is generally recognised for avoiding "over-fitting", controlling sampling error and performing better validation. Many of the arguments against large samples fail adequately to recognise the so-called "curse of dimensionality", a phrase used to highlight the fact that when large numbers of variables (dimensions) are being considered, every point becomes a statistical outlier.

We shall conclude that in order to develop and understand more powerful, integrated scoring models, it is highly beneficial to take advantage of the ability of modern computing technology to handle much larger datasets than are conventionally used.

1 Introduction and Motivation

Current hardware and software allows datasets with hundreds of thousands or even millions of records to be handled with comparative ease, and yet most scorecards are built with training sets[1] of only a few thousand records, with perhaps 5,000–20,000 being the most common size. This paper looks at the potential advantages of using significantly larger training sets, and the dangers of failing to do so.

Throughout this paper, much of the emphasis is on the importance of non-additive effects even in the context of conventional additive scoring, and much of the motivation for larger sample sizes will derive ultimately from the “curse of dimensionality” (Bellman, 1961).

We begin in section 2 with some extremely simple illustrations of the problems of using small datasets by looking at the variation in bad rates as a function of randomly generated variables to give a feel for the scale of the problem, and to illustrate the basic ideas of interactions and the curse of dimensionality. This section is largely descriptive and obvious, and may be skipped by anyone who simply wants to “look at the numbers”.

After this, in section 3 we will present some straightforward experiments with varying sample size in the context of additive scorecard development without interaction variables or segmentation of the dataset. The results indicate that even in this simplest of cases, there are benefits to be derived from significantly larger datasets than are used conventionally, clearly indicating significantly greater benefits for more complex models.

In section 4, we assume that some given interactions (non-additive effects) are known to exist, and explore the implication of a desire to exploit these on dataset size. Here, the assumption is that we do not need to *detect* the interactions (because they are given), but merely to develop scorecards that exploit them, either by introducing interaction variables, or by segmenting the population. Here we shall see that rather larger sample sizes are required for training than is the case when no non-additive effects are to be catered for.

Section 5 considers the case in which it is assumed that some interactions may exist, but that they have not been identified. In this case, the problem becomes both to identify and exploit the interactions. We consider a variety of means of detecting the interactions, and argue that still larger training sets are needed in order to make this process effective.

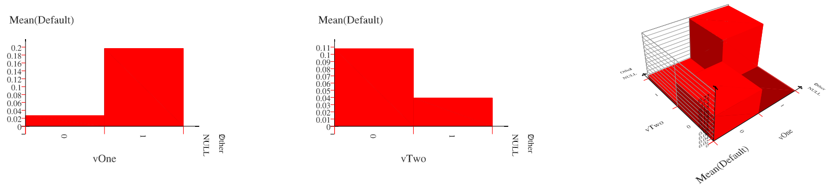
2 Sampling and the Curse of Dimensionality

At base, the reason we shall find that we can benefit from large samples is that we are interested not only in single variable distributions, but distributions of two or more variables (cross-distributions). For example, we shall be interested not only in bad rate as a function of age, and bad rate as a function of income, but also of bad rate as a function of age *and* income.

2.1 Interactions

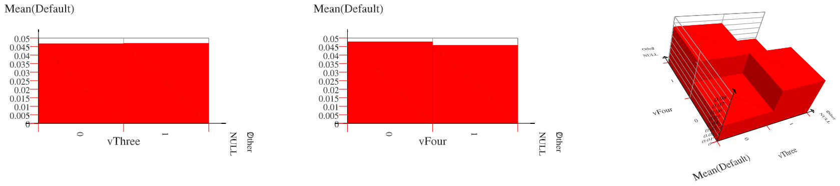
Figure 1 shows one case in which we might be interested in the considering two variables at once. The graph on the left shows the bad rate as a function of some binary variable x_1, showing a much higher bad rate for customers with $x_1 = 1$ than for those with $x_1 = 0$. The centre graph shows a similar picture for a second variable x_2, except in this case customers with $x_2 = 1$ exhibit a significantly lower bad rate than those with $x_2 = 0$. From these simple plots, one might naïvely assume that customers with $x_1 = 1$ and $x_2 = 0$ (the two “bad” values) would have an even higher bad rate, but in fact the plot on the right, showing bad rate as a function of the four

[1] We shall call the historical data available during model building the training set and the data used to assess the performance of the model the test set.



Bad rate $f(x_1)$ Bad rate $f(x_2)$ Bad rate $f(x_1, x_2)$

Figure 1: The single-variable bad rates show that $x_1 = 1$ is correlated with bad debt (left plot), as is $x_2 = 0$ (centre plot). We might therefore expect that the having both of these attributes would lead to a very high bad debt rate, but in fact the right-hand plot shows that there is a rather low bad debt rate associated with this group. This is an example of a non-linear effect, giving rise to non-additivity.



Bad rate $f(x_3)$ Bad rate $f(x_4)$ Bad rate $f(x_3, x_4)$

Figure 2: In this case, the bad rate is largely independent of the single variable values x_3 and x_4, as can be seen from the flat profile in the left and centre graphs. It can be seen, however, that when both variables are considered, there is considerable dependence on the values of x_3 and x_4. This is another example of a non-linear effect, again giving rise to non-additivity.

possible pairs of values for x_1 and x_2, it can be seen that the bad rate is actually rather low for people in this group.

The situation illustrated is an example of a *non-additive* effect, also known as an *interaction*. Such effects are largely incapable of being modelled by simple additive scorecards when the characteristics are identified with the variables x_1 and x_2, because the score contribution assigned for having $x_1 = 1$ is independent of the value of x_2. In this case, depending on the relative populations of the four groups, it is likely that both variables would be included in the scoring model with reasonable weighting, but that the interaction between the variables would degrade the performance of the scorecard. Although a battery of techniques for accommodating non-additive effects has been developed and is used in the credit scoring community, and these are discussed in later sections, we think it is fair to say that non-additivity can be a major source of inaccuracy in this type of scoring model.

Figure 2 illustrates a similar, but subtly different non-additive effect. In this case, neither x_3 nor x_4 is much use as a predictor independently, but the dependence of bad rate on their joint distribution can be seen from the graph on the right to be considerable. In this case, a simple analysis would lead to the exclusion of these variables from the scorecard, whereas in fact it can be seen that they have considerable predictive utility.

2.2 Sampling Cross-distributions

The principal techniques used within the scorecarding community to handle non-additive effects are the creation of interaction variables and segmentation of the population into groups within which interactive effects are relatively small. We consider both in detail in later sections. Here, we focus particularly on the considerations that apply when sampling over more than one variable, which is a necessary component of both approaches.

2.2.1 Some Very Simple Graphs

Suppose we have a population of 100,000, and two variables (characteristics) each of which we divide into ten bins (attributes, or bands). If we use "equal population" bins (i.e. choose the bin boundaries so that roughly equal numbers of records fall into each bin for each variable), it is clear that we will have about 10,000 in each bin for each variable, and equally clear that if we consider two-variable bins, we will have only around 1,000 records per bin. This is illustrated in the top line of figure 3. In this case, both characteristics (x_1 and x_2) are random variables in the range $[0, 1]$, generated from a uniform distribution using the Marsaglia pseudo-random number generator (Marsaglia, 1984; Marsaglia *et al.*, 1990). An immediate point to notice is that—naturally—not only are the bins for the cross-distribution smaller, but the proportionate variance in their size is much larger. This is simply one manifestation of the "curse of dimensionality" (Bellman, 1961, Friedman, 1997), illustrating that as the number of dimensions grows (in this case just from one to two!) any sample becomes increasingly poor.

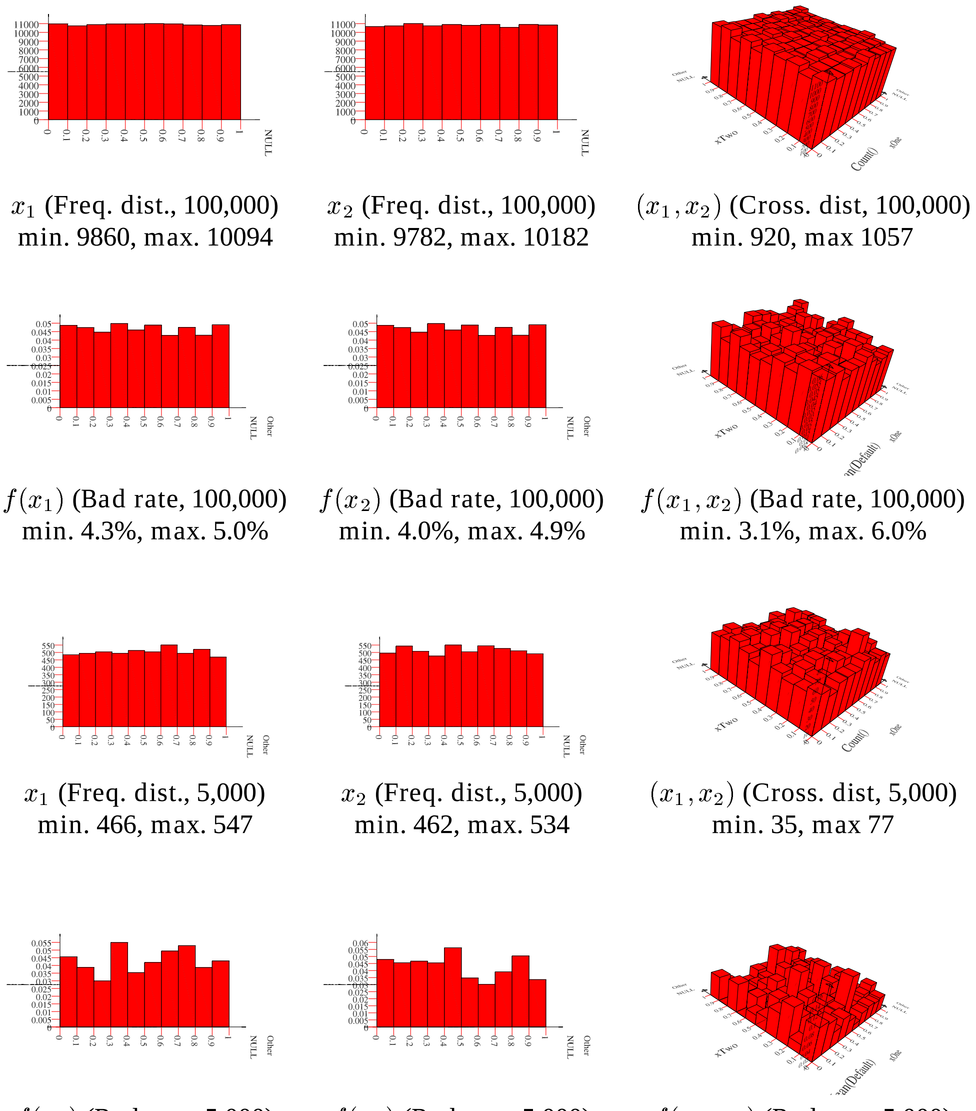
The second line of figure 3 shows what happens when we now consider bad rate as a function of these two variables, both independently (left and centre) and together (right). It is important to remember here that x_1 and x_2 are uniform random variables. The bad rate is computed simply as the ratio of the number of "bads" in the bin (n_b) to the total number of records in the bin ($n_b + n_g$). Notice how the bad rate varies from 3.1% to 6.0% in the joint bins, even though there can be no true dependence between these variables at all, and how much larger this variation is than the corresponding variation in the single-variable bins. (The actual bad rate over the 100,000 records is 4.9%.)

The third and fourth lines of figure 3 correspond exactly to the first and second respectively, except that they now illustrate the corresponding situation for a (single) uniform random sample of 5,000 records from the 100,000. The key point to note are that all the variances are significantly increased, especially in the cross-distributions. This is, of course, simple a function of the reduced effective sample sizes in each bin. (The overall bad rate in the sample of 5,000 is 4.6%.)

2.2.2 Discussion of the Simple Graphs

In scorecarding, the fundamental quantity that we need to estimate on a bin-by-bin basis is the bad rate $n_g/(n_g + n_b)$. The graphs presented in the previous section suggest that we can quite easily form rather poor estimates of these quantities, and that the likelihood of this increases with decreasing overall sample size, and with decreasing bin size. When considering two or more variables together, the dangers of small samples evidently increase exponentially. However, a number of objections may be raised to this broad line of argument, and we now discuss these in turn.

- *"You don't need more goods than bads."*
  It is often argued that if only a small number of "bads" are available, there is little benefit in using more "goods", and indeed it is sometimes argued that it is actually harmful to use a very unbalanced sample with (say) all the "goods" as well as all the "bads". From the point of view of estimating the quantity of principal interest to us—the bad rate—this is quite wrong. The bad rate $n_g/(n_g + n_b)$ depends on two quantities, n_g and n_b, and improving the accuracy with which either is known decreases the error associated with the bad rate. There are, however, implementational reasons why this benefit may in practice be outweighed by other factors, and these are discussed in section 3.1.



x_1 (Freq. dist., 100,000)
min. 9860, max. 10094

x_2 (Freq. dist., 100,000)
min. 9782, max. 10182

(x_1, x_2) (Cross. dist, 100,000)
min. 920, max 1057

$f(x_1)$ (Bad rate, 100,000)
min. 4.3%, max. 5.0%

$f(x_2)$ (Bad rate, 100,000)
min. 4.0%, max. 4.9%

$f(x_1, x_2)$ (Bad rate, 100,000)
min. 3.1%, max. 6.0%

x_1 (Freq. dist., 5,000)
min. 466, max. 547

x_2 (Freq. dist., 5,000)
min. 462, max. 534

(x_1, x_2) (Cross. dist, 5,000)
min. 35, max 77

$f(x_1)$ (Bad rate, 5,000)
min. 3.0%, max. 5.5%

$f(x_2)$ (Bad rate, 5,000)
min. 3.0%, max. 5.6%

$f(x_1, x_2)$ (Bad rate, 5,000)
min. 0.0%, max. 13.0%

Figure 3: The top line shows the frequency distribution of two uniform random variables (characteristics) divided into ten bins (attributes) and their cross-distribution on a dataset of 100,000 records. Note that (naturally) the bin size is a factor of ten smaller on the cross-distribution, and the variance in bin size is proportionately much higher. The second line shows the bad rate as a function of these variables. Since the variables are randomly generated, there is no real correlation with the bad rate. The lower two lines are the same as the top two except that they show the distributions and bad rates for a random sample of 5000 records from the 100,000. Notice the much larger variations in bad rate both for the two random variables independently and (even more) for their cross-product.

- *"It's not a problem if you use stratified (grid) sampling."*
  It may objected that the preceding discussion uses non-stratified samples, whereas a good sampling procedure is stratified. This is true, and indeed is also discussed in section 3.1. However, there are severe problems with stratification in large numbers of dimensions. It is true that we could stratify by at least determining the proportion of "goods" and "bads" that we sample, and there is every reason to do so. If we did this, the variances seen in the various graphs would diminish. However, it is infeasible to stratify the entire sample, and this is again due to the curse of dimensionality.

  Consider even a very modest case of a scorecard with ten variables, each with five bins. This gives 5^{10} cells to consider (a little under 10 million). In principle, we would like enough points in each cell to form a reliable estimate of the bad rate! Such fully stratified (or "grid") sampling is clearly infeasible for any normally complex problem, even given an unlimited universe of records from which to draw. In practice, the situation is even worse, because many of the combinations of attributes will not even exist in the population. This is the essence of the curse of dimensionality, and justifies the claim that in high dimensional space, 'all samples are poor samples'.

- *"Interactions don't matter for additive scorecards."*
  A fundamental property of additive scorecards is that they are, as the name suggests, additive, i.e. the score contribution from each variable (characteristic) is independent of the values of other variables. It can therefore be argued that you only need to get a sample size good enough to represent each bin (attribute) properly for the variable with most bins, and it will be a good enough sample overall. There is a good deal of truth in this, and our main motivations come from consideration of handling interactions directly (sections 4 and 5), but there are also some subtleties even in the case where interactions are not being directly considered. First of all, we are *not* performing a set of independent single-variable (binwise) regressions and then combining the model using Bayes Theorem (e.g. Hand, 1981) or some other combining methodology (e.g. Breiman, 1997). Rather, we are fitting a set of hyper-surfaces with as many degrees of freedom as there are bins in all the variables, and thus all variables do contribute to the regressions performed. This often results in the weights for individual variables being significantly different from those that would result from single-variable regressions, because of the non-uniform distribution of samples throughout the sample space. Secondly, note that we are not estimating the bin-wise bad rates for each variable independently—we are using exactly the same set of records to estimate the rates for each variable, potentially leading to artificial correlations in the errors. Thus, although the curse of dimensionality is less problematical for additive models than for other types of models, it does still have a (negative) effect.

2.3 Practical Limitations

To recapitulate, we have argued that

1. our key task is to estimate the bad rate for various customer segments (bins or intersections of bins);
2. in principle, larger sample sizes can only help with this task;
3. when we move beyond single-variable models, large samples are required to develop reasonable estimates of $n_b/(n_b + n_g)$.

In order for increasing sample size to be assured (statistically) of having this beneficial effect in practice, three conditions must be met:

1. the quality measure that we use to assess scorecard performance must be insensitive to the ratio of "goods" to "bads" in the population. More precisely, if we replicate any set of records, the performance measure should not change. For example, if we took all the "goods" and duplicated them, this should not change our performance measure.

2. the solver (optimiser) must be optimising the given quality measure directly, not some surrogate function.
3. the solver (optimiser) must optimise the model correctly (i.e. find the global optimum of the quality measure).

These conditions, which are discussed in more detail in Radcliffe & Surry (1997), are *not* usually met in practice. For example, consider the "standard" logistic regression scorecard measured with Gini. The quality measure, GINI, actually does satisfy the stated condition, because it depends only on the *ordering* assigned to records by the scoring model. However, when developing a logistic regression model, the quality measure actually used to perform the fitting is a fitting error (to be minimised), and this is strongly affected by every point in the dataset. If the good records are doubled, the model will be (loosely) twice as interested in fitting "goods" as when each is present only once. Thus the quality measure *seen* by the optimisation is not the one standardly used to assess the output of the scorecard build. Finally, unlike "linear" scorecard models, logistic models cannot be solved directly, so an indirect, iterative solver (such as Newton-Raphson, or conjugate gradient) must be used, and such solvers do not, in general find a global optimum of the function to which they are applied.

What this means in practice is that a larger development sample can result in worse (observed) performance of the scoring model, even though the estimates of the bad rates in each segment considered should be better. However, it is possible to avoid most of these problems by careful manipulation of the form of the objective actually seen by the solver, and taking care over the regression itself.

3 Single Additive Scorecards without Interaction Variables

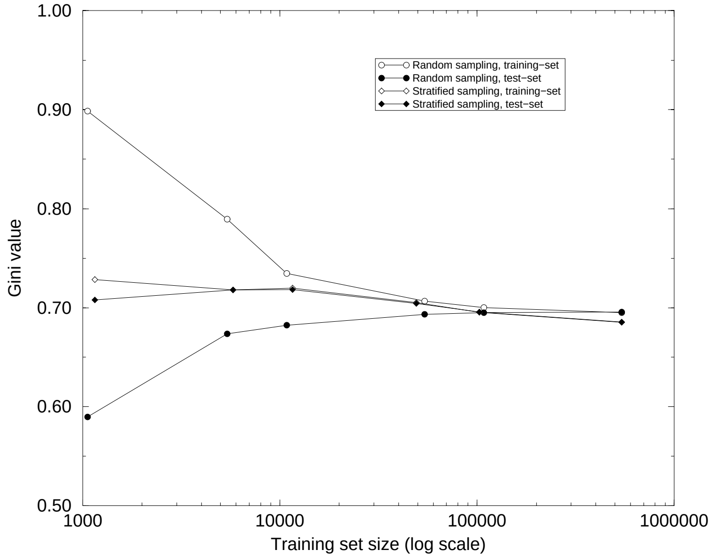
3.1 Random and Stratified Sampling for Training Sets

Our first experimental results concern a relatively common case, namely additive scorecards without interaction variables. We start from a historical sample of around 1 million records, with a bad rate of around 1%, so that we have 10,000 "bads" available. We consider a fixed set of ten characteristics, each with ten equal-population attributes. We then repeatedly build scorecards using (local, bin-wise) linear regression models with a test set of roughly constant size (approximately half a million records) but vary the training set set from 5,000 to 500,000. The local regressions are solved directly and precisely using a maximum likelihood assumption, which effectively minimises the fitting error, and it should be noted that this measure is *not* independent of the bad density, so we might expect some problems when the ratio of "goods" to "bads" is not close to 1. We repeat these experiments for both weights-of-evidence models and non-weights-of-evidence ("dummy variables") models, and in each of these cases use both random samples and samples balanced so as to have either a (roughly) equal number of "goods" and "bads", or (for the larger sample sizes) all the "bads" and the remainder of "goods". No other form of stratification was applied.

The results of these experiments are shown in figure 4. These results illustrate many of the problems anticipated above, showing that when we optimise against a function that is sensitive to the density of "bads" in the sample, performance actually decreases with increasing sample size.

3.2 Reweighting Bads

To mitigate these effects, we repeated the experiments with models in which the bad records were weighted in such a manner as to negate precisely the effect of density variation. The results for these experiments are shown in figure 5. In these cases, as would be expected, performance continues to increase slightly with increasing sample size, and even the point where



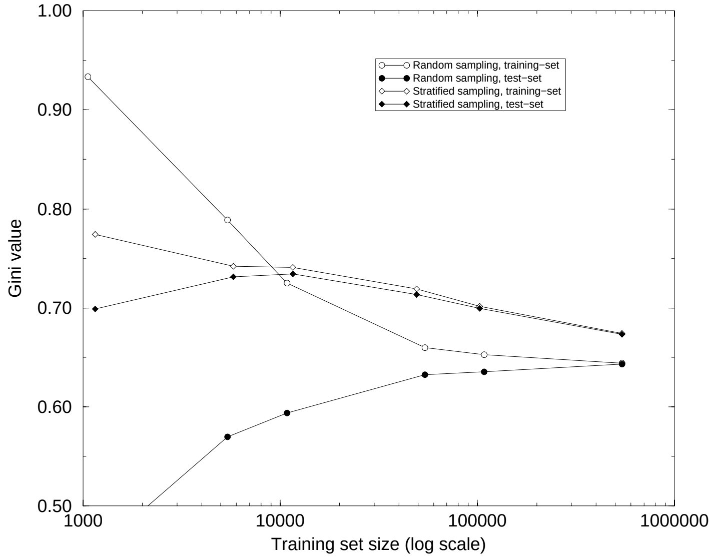


Figure 4: These graphs show the average Gini over ten runs on both the test set (filled symbols) and the training set (open symbols) for a variety of sizes of test set. In each case a test set of around half the population (c. 500,000 records) was reserved and training sets were drawn as samples from the remaining 500,000 records. The top graph shows performance using a one-weight-per-characteristic (weights-of-evidence) model, using both a training set that is a pure random sample from the population, and one which is stratified so that there are a roughly equal number of “goods” and “bads” for small test sets, and in the case of the larger test sets, so that all the available “bads” are included and the remainder of the sample is made up of randomly selected “goods”. The bottom graph shows corresponding results for a one weight per attribute (“dummy variables”) model, with both pure and stratified random samples. Clearly performance on the test set peaks at relatively low training set sizes when using stratified sampling, in the range 5,000–10,000, and performance declines thereafter. This is due to the form of fitting error used in the regression, as is discussed in the text.

strongly diminishing returns sets in is raised to training set sizes of at least 100,000. Thus these results provide modest encouragement to us in our intuitive belief that the more accurate estimates of the bad rate $n_b/(n_b+n_g)$ should lead to increasing performance, albeit with asymptotic behaviour for large sample sizes.

4 Scorecards with Interaction: Non-linearity and the Curse of Dimensionality

In the previous section we have seen evidence that provided proper care is taken, even conventional additive scorecards can benefit from large sample sizes. We now consider the benefits available from including interaction effects in scoring models. There are two main ways to achieve this. The first is to build "interaction variables" which are single new variables that summarise the effects of two others. For example, referring back to figure 1, we could build a four-valued variable that summarises the four possible combinations of values. (This is a special case in which the new variable has the same number of bins as the total number of bins in the two source variables, because $2^2 = 2 + 2$, but in the general case the natural interaction variable has more bins.) The other principal method for handling interactions is to segment the population on one or more of the variables involved in interactions and to build separate scorecards for each subpopulation. The current section is not concerned with the problem of *detecting* interactions, but simply with exploiting them; the detection of interactions is discussed in the next section.

Suppose, then, that we know by whatever means that two given variables interact, and that we form an interaction variable from these. Here we are interested in the benefits available from exploiting the interactions by including the interaction variable in our model, and the sample sizes required to do so.

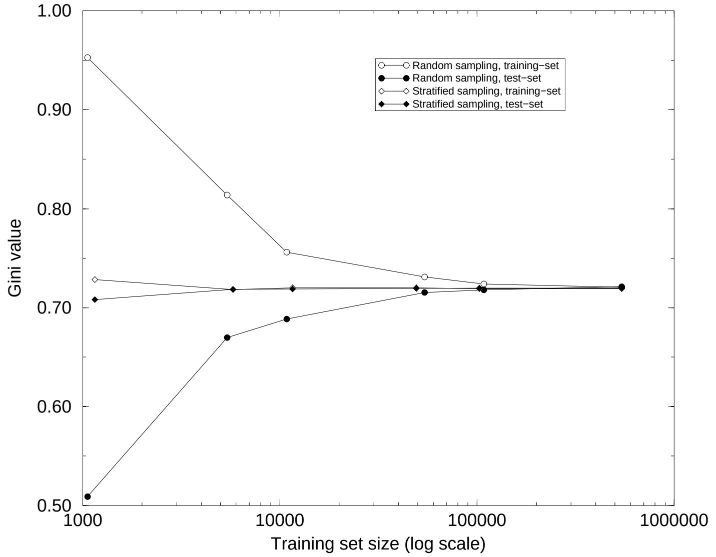
Once we have created the summary variable, it could be seen conceptually as "just another characteristic", no different from any other in the scorecard. Thus it might be argued that the sampling requirements are no different from any other variable. However, this misses the mark in two important respects. First, precisely because it is an interaction variable, it is likely to be highly correlated with existing characteristics, and thus affected by the kind of systematic sampling biases discussed above. More importantly though, the new variable typically divides the population into highly non-uniform segments. This is because the interaction tends to capture a weakness in a simpler model over one or more relatively small segments; giving a relatively large improvement in those segments (consider again figures 1 and 2). Thus the sample sizes required to achieve robust estimates across the range of the interaction variable are necessarily increased, compared to those required for characteristics which partition the population into more or less equal sized segments.

Similar arguments apply when an interaction is used to segment the training data set, with a separate additive scorecard being constructed for each. Here we clearly need sufficient data volumes to achieve accurate estimates across all subpopulations.

Initial quantitative investigations of both these effects show promising results, and systematic experiments are underway.

5 Interaction Variables, Segmentation and Variable Selection

The final aspect that we consider, having established in the previous section that interactions can be important, and that modelling them can pay significant dividends, is to consider how we might detect interactions, and thus find either useful segmentations or useful interaction variables. It seems clear that in this context the curse of dimensionality and the problems of small



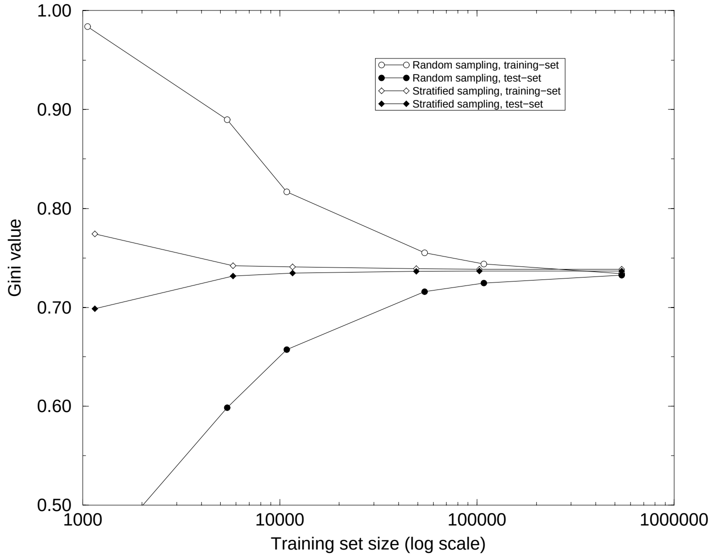


Figure 5: These graphs show the the same experiments as in figure 4 except that now the "bads" have been weighted in such a way as to give them the same weight as the "goods" in computing the fitting error for the regression, even when there are significantly fewer of them. As intuition would suggest, in this case, performance does *not* deteriorate with increasing training set size, and indeed, continues to improve certainly up to a sample size of 100,000, and arguably to 500,000, although clearly the benefits are modest.

samples are going to be more acute than when simply exploiting interactions already known, so we should expect more benefit from larger sample sizes.

We can imagine, and have investigated, various methods for detecting interaction variables. Once particularly attractive method is to model the error on the training data (or, with great care, the test data) using non-additive (non-linear) modelling techniques. That is, we compare the model's prediction to the known outcome, and attempt to characterise those segments on which it performs poorly. This could be done either by choosing a cutoff score and modelling false positives or false negatives together or separately, or by modelling the (continuous) difference between the predicted default probability assigned to each record and its actual (defaulting or non-defaulting) historical behaviour. Perhaps the most obvious such technique is some form of decision tree, such as CART (Breiman *et al.*, 1984), especially if some form of lookahead is used rather than the usual one-step optimal approach. Combinations of variables that show up strongly in the tree are good candidates for interaction variables, and single variables appearing in the tree, especially towards the top are good candidates for segmentation criteria.

Because we are now seeking non-linear interactions by examining multi-dimensional relationships, all of the arguments in section 2 apply, and suggest the need for significantly larger sample sizes than those required in the simple situation of section 3. Effectively, we no longer desire simply to estimate the bad rate within a fixed segment of the population, but are attempting to define segments that have significantly "different" rates from those which the model predicts. For example, in the example of figure 3, a robust model using only the two variables x_1 and x_2 would (hopefully!) predict an equal default rate across all variable values. Clearly no interaction between the variables exists in reality, but consider the potential for misinterpretation if we searched in the (x_1, x_2) space for highly misclassified regions (bottom right diagram in figure 3).

As the methodology for searching for new interaction variables is difficult to completely automate, it is somewhat difficult to define a "fair" experiment to illustrate these effects. However, we have successfully used the methodology in practice and are working towards quantifying the effects more rigourously.

6 Summary

By building and testing several hundred scorecards on sample sizes of up to half a million records, we have demonstrated that even in the simplest case signficant benefits can result from using substantially larger samples than has traditionally been the case, and in fact larger than the biggest possible balanced sample (taking all of the "bads" and the same number of "goods" at random).

At its most basic, the key is to make best estimates of the bad rate within each attribute grouping. When characteristics and attributes are fixed, the advantages of larger samples rest on well-known statistical arguments. Of more interest in the incorporation of interactions in the standard additive scorecard. This requires both that relevant interactions be *found*, by examining multidimensional distributions, and that they be weighted properly within the model. Both of these tasks place stronger requirements on the samples sizes necessary to achieve the accuracy required to develop robust models with respect to test-set data; ongoing investigations seek to further quantify these results.

References

R. E. Bellman, 1961. *Adaptive Control Processes.* Princeton University Press.

L. Breiman, J. Freidman, R.A.Olshen, and C. Stone, 1984. *Classification and Regression Trees.* Wadsworth.

L. Breiman, 1997. Arcing classifiers. Technical report, Statistics Department, University of California at Berkley.

J. H. Friedman, 1997. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Knowledge Discovery and Data Mining*, 1(1):55–77.

D. J. Hand, 1981. *Discrimination and Classificatrion*. John Wiley (Chichester).

G. Marsaglia, A. Zaman, and W. W. Tsang, 1990. Toward a universal random number generator. *Statistics and Probability Letters*, 9(1):35–39.

G. Marsaglia, 1984. A current view of random number generators. In *Computer Science and Statistics: 16th Symposium on the Interface*.

N. J. Radcliffe and P. D. Surry, 1997. Credit scoring as an optimisation process. In *Proceedings of Credit Scoring and Credit Control V*. to appear.