Measuring Confidence Levels in the Predictions of Credit Risk Models.

Authors: Mark Stirling, Paul Robinson
Paragon Business Solutions Ltd
e-mail: mark@credit-scoring.co.uk

Abstract

Credit scoring regression models are routinely used in binary decisions where a cut-off score indicates a pass/fail threshold. Increasingly, the predicted values from these models are used to drive other decisions and calculations including loan pricing, product downsell strategies, behavioural marketing strategies, collection activity and not least, predicted losses in Basel type calculations. In most implementations, little regard is given to the error level of the underlying score and hence calculations and business decisions in which the score is used. This is especially prevalent in respect to the change of error level at different points across the score distribution.

We propose a technique to derive optimal score bands to minimise the experimental and theoretical variation in predictions. We will show how to measure prediction confidence levels and observe their variation across the full range of scores using non-parametric techniques and we will demonstrate the advantage of these techniques over Standard Error calculations.

Introduction

In modern business practice, much reliance is based on regression models to predict (usually) probability of credit default, both at application time and then during the life of the credit relationship.

When scorecards are used in a classical accept / reject type decision, the available population is split in two, resulting in large sub-sets of data. The scores obtained may then be used to predict results (bad rate) for each sub-set. The relatively large sub-sets ensure that the variance in results is therefore correspondingly small.
However, if the same scorecard is used for spot or marginal decisions, for example when an individual risk prediction is used in a further calculation such as in Risk Based Pricing or in calculations of customer value or expected loss, the confidence in results is typically not tested.

This problem is addressed to some extent in model building when standard validation techniques measure actual vs expected results within the development data and, more relevantly, if a smaller out-of- sample (holdout) set is used to measure results. However, these techniques are usually applied judgmentally and any local variation is typically written-off as “sample size problems” so that the predicted values may be applied as fact.

To reliably use the prediction results from a scorecard in a well-focussed or marginal area, we need to arrive at some measurement of confidence for the predictions which

is stated over the whole range of possible scores and is either a blanket measurement or, more likely follows some trend over the score range.

Possible Solutions

A number of tools from the standard toolkit of scorecard developers operate in this area. They may serve their stated purpose very well, but are of limited utility in addressing the confidence level problem:

Testing the Actual vs Predicted Results:
A result of using logistic regression techniques is that the actual results should match the predicted results over samples of the data. This test is usually performed as model validation against a score distribution and against the attributes of the scored characteristics.
Problems with this technique are :

- Measurement over the score distribution is prone to sample size problems due to inappropriate score bands.
- The lack of an out-of-sample test may lead to over-fitting problems.
- Assessment of the fit is usually judgmental based on graphical results which may be influenced by scaling etc.
- The single sample does not lend itself to measurement of confidence boundaries.

Holdout Sample Comparison:
Typically, a sample of the development data is held back for validation purposes so that actual results may be compared to the development data.
Problems with this technique are :

- Measurement over the score distribution is prone to sample size problems due to inappropriate score bands. If the bands are adjusted to cater for the (usually) smaller sample size in the holdout, the test cannot be truly independent.
- Assessment of the fit is usually judgmental based on graphical results which may be influenced by scaling etc.
- The single sample does not lend itself to measurement of confidence boundaries.

New Data Samples:
Often a new data sample will be obtained for validation of the scorecard, this will usually consist of a newer sample of data which is immature with respect to the scored outcome period.
Problems with this technique are :

- An immature sample will tend to demonstrate fewer “bad” cases and this must affect any measurement of actual vs expected results.
- The same problem of score banding exists.
- Assessment of the fit is usually judgmental based on graphical results which may be influenced by scaling etc.
- The single sample does not lend itself to measurement of confidence boundaries.

Measuring Confidence Levels in the Predictions of Credit Risk Models.

Standard Error Calculation:

$$SE = \frac{\hat{\sigma}}{\sqrt{n}}$$

where:

$\hat{\sigma}$ is an estimate of the standard deviation σ of the population
n is the size of the sample

Standard error will produce a statistical measure of confidence levels around our predicted values. At first glance, this would appear to meet our stated requirements. However, Standard Error calculations assume a normal distribution of errors around the measured value. This is implied by the Central Limit Theorem as sample sizes tend towards infinity, but since we are sub dividing our population into score bands to produce confidence intervals at different points on the score distribution, our sample sizes will become increasingly smaller and this assumption may not hold.

With this caveat, Standard Error may be estimated for score-bands as :

$$SE = \sqrt{\frac{BR_{Band} \times GR_{Band}}{n}}$$

where

accounts are divided into Good and Bad
BR_{Band} is the percent measurement of Bad accounts in a band
GR_{Band} is the percent measurement of Good accounts in a band

The Proposed Solution

Whereas the ideal solution would be to state confidence levels as a function of the score, or to measure the levels for individual scores, the problem of diminishing sample sizes dictates that we can only measure error rates on score bands. It should be obvious then that the size of score band must be optimised for confidence measurement. If a score band is too small, variation in the predicted values will adversely affect the confidence of our measurements. If a score band is too large, a similar problem exists as any given score can only be said to exist at the mean point of the band for predictive purposes.
It should also follow that score bands should not necessarily be constant size over the score distribution. Indeed, we would expect to see wider bands at the extremes of score where fewer data samples exist.
Our first requirement therefore is to produce an optimised set of bands for the measurement of predicted values.

Having achieved a suitable set of bands, we then require a number of measurements of the bad rates achieved in those bands for different data samples. We will use bootstrapping to simulate a large volume of different data samples. The maximum and

minimum predicted values from the bootstraps then represent the extremes of our confidence range.

Bootstrapping

Bootstrapping is the method of resampling a data set *with replacement* such that summary statistics may be estimated over many different measurements.
Bootstrapping is particularly fit for our purpose as:

* It does not require the normality assumption to be met. The results may not be normal or symmetric.
* It can be used with small sample sizes which may be necessary when considering small score bands.

The Banding Problem

There appear to be two options for the use of bootstrapping in optimising score bands. Bootstrapping may be applied to the entire population, resulting in variance in the volume of records falling into any given scoreband over the samples, or the initial samples may be fixed according to the development population for every possible band to be considered and bootstraps then created for each (potential) band individually.
While undoubtedly computationally expensive, the second option gives the benefit of constant volume within a band across the bootstrap samples, therefore simplifying calculations. However, this would in effect constrain the volume distribution by score to mirror the development population at a macro level (dependant on band size) and we have therefore chosen to use the whole sample as our bootstrap base.

Given the possible variation of population volume across any given score band, we can see that the mean predicted bad rate for the band may vary anywhere between the maximum and minimum boundaries.

e.g. consider the band consisting of the score range from 101-105 (in the logistic model using 20 points to double the odds and odds of 1:1 at a score of 100).

Development Sample		
Score	Volume	Predicted
		Prob(Good)
101	12	0.51
102	10	0.52
103	11	0.53
104	12	0.53
105	10	0.54
Mean P(G)		0.53

Bootstrap 1		
Score	Volume	Predicted
		Prob(Good)
101	55	0.51
102	0	0.52
103	0	0.53
104	0	0.53
105	0	0.54
Mean P(G)		0.51

Bootstrap 2		
Score	Volume	Predicted
		Prob(Good)
101	0	0.51
102	0	0.52
103	0	0.53
104	0	0.53
105	55	0.54
Mean P(G)		0.54

Since the ratio of Good to Bad accounts in the bootstrap samples will vary, we can also measure the actual Probability(Good) for each sample. We can take the distribution of these samples to provide a measure of actual range to compare to the predicted range. The boundaries of our score band may then be expanded and

contracted to provide different measurements. The optimal band will be said to exist where the actual range of predicted values matches the predicted range.
In practice the actual values will vary between all good (100%) and all bad (0%) and it is necessary to top & tail at some level to

e.g. consider the following possible score bands starting at score 101. The optimal band size is 6 points. i.e. 101-107 where the predicted P(G) range matches the bootstrap range. Any of the potential ranges smaller than this show the bootstrap range larger than the predicted range and the situation is reversed for larger ranges.

	Predicted		Actual (Bootstrap)	
Band Size	Min P(G)	Max (PG)	Min P(G)	Max (PG)
2	0.51	0.52	0.20	0.8
3	0.51	0.53	0.40	0.6
4	0.51	0.53	0.50	0.55
5	0.51	0.54	0.51	0.55
6	0.51	0.55	0.51	0.55
7	0.51	0.56	0.50	0.55

Further consideration should then be given to adjacent bands. Different results may be obtained dependant on the start position of the first band. E.g. in the above example, we may decide to make a band from 101 to 107. The start point of the next band would then be 108. However, the 101 point was an arbitrary choice and it would be equally valid to start from the highest obtainable score and work downwards.
The effect of different start points or other banding techniques is thought to be minimal and we have chosen not to investigate further.
In practice we have calculated, for each integer score, the optimal band size using that score as a start point. We have then used these values to assign score bands on a contiguous low score to high score basis.

Score	Optimal Band	Assigned Band
101	5	A
102	5	A
103	6	A
104	7	A
105	7	A
106	7	B
107	6	B
108	5	B
109	6	B
110	5	B
111	5	B
112	4	B
113	3	C
114	3	C
115	3	C
116	3	D
117	3	D
118	3	D
119	5	E
120	5	E
121	5	E
122	7	E
123	5	E
124	4	F
125	3	F

Measuring Confidence Levels in the Predictions of Credit Risk Models.

Practical

Following on from the principals and theory laid out in the previous section, the second part of this paper deals with a worked example. This section includes an overview of the dataset, and notes on a scorecard produced on this data, the optimisation of the banding for this data, producing the confidence levels and a discussion on the relevance of the results for implementation and practice.

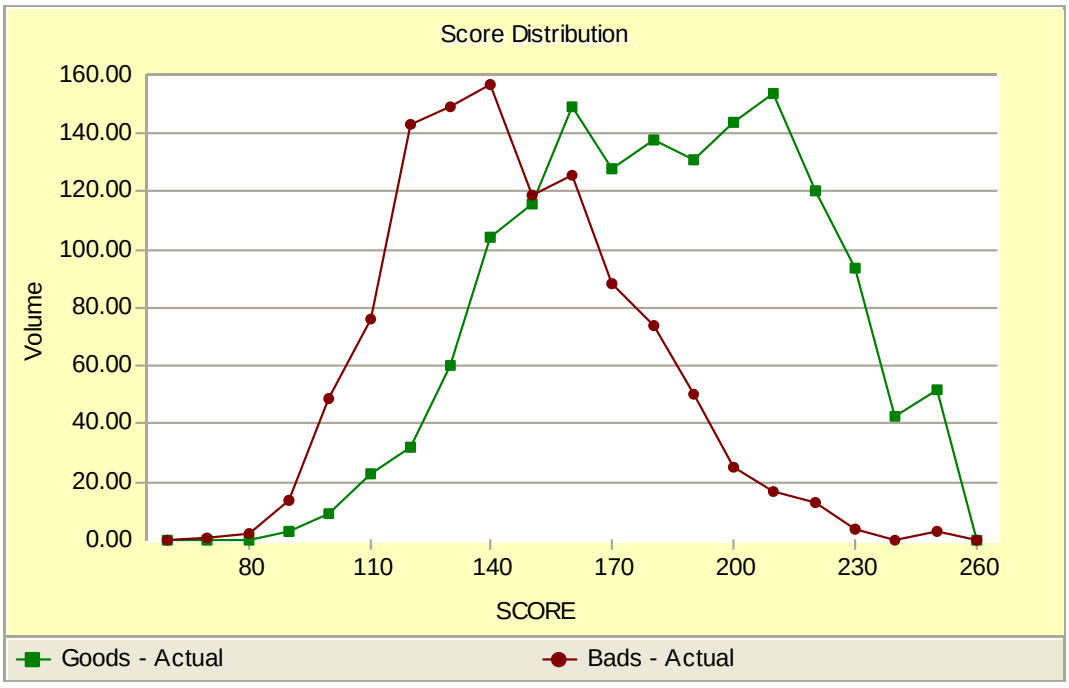
The data

This worked example uses a small credit application dataset from the 1990's consisting of 1500 Goods and 1100 Bads. The dataset only includes accepted records and the bad definition is a standard 3+ payments in arrears.

The scorecard

The scorecard which has been produced using this data is a standard, logistic, weight of evidence model formed using forward stepwise selection of characteristics. Characteristics included are: Customer Age, Time with Bank, Employment Term, Housing Status, Home Term, Bureau Searches, Number of Credit Cards Held, Telephone Ownership.

This scorecard separates the good and bad accounts by score, with a Gini of 58 and the following score distribution:



Measuring Confidence Levels in the Predictions of Credit Risk Models.

Band Optimisation

As discussed above the optimal bands must be obtained prior to carrying out the error analysis. This ensures that the results produced can be relied upon and compared with other results.

This band optimisation uses the method described earlier which produces a band whose bootstrap samples contain the same range of bad rates as the predicted score range.

When carrying out this work in practice certain tolerances should be applied to the results, given that we are constrained to operating with integer scores, and definitive values.

When performing the bootstraps a top and tailing mechanism will reduce the bias towards results featuring 0% and 100% bad rates. The higher the top and tailing percentage, the smaller the bands will be that are produced. It is important that this value remains small, so as not inadvertently effect the resulting optimisation. For this worked example a top and tailing value of 10% was chosen.

Secondly it is unlikely that the range of actual bad rates and the predicted score range bad rates are to exactly equate. Therefore a tolerance in the match of these figures has been devised. In this worked example a tolerance of ±10%. i.e. if the predicted bad rate range is within 10% of the actual bad rate range from the bootstraps then the ranges are said to equate.

Two other variants are required to be determined. Firstly the number of bootstraps samples to take. The bootstrap samples reflect a range of possible outcomes, given the original distribution. If only a few bootstraps are taken, say, 20, then the number of different combinations of the original data is going to be limited by the number of bootstrap samples taken. What is the correct number of bootstraps to use? The number of samples created should be larger enough not to limit the possible combinations produced by re-sampling. As the number of combinations is particularly large (as determined by binomial counting theorem) it is suggested that the largest constraint on the number of bootstraps is performance related. In this worked example 1000 bootstraps samples have been produced.

The second variant is which range of bandwidths should be considered? Quick experimentation can show that the smallest bandwidths are unlikely to be involved (1-5) as the central points of the distribution are generally shown to be represented in bandwidths of greater than five. However this is now determined primarily by experiment (due to the experimental/randomisation of the bootstrap sampling technique). At the very edges of the distribution (where concern is less likely to be raised upon the banding) it may be possible that very large bands are determined. Again by experimentation this has proved to be in the order 15-20 point bands. At the tails of the volume distribution it may be that a bandwidth of 20 is insufficient, in which case the bands will be expanded to incorporate all values – thus creating the highest and/or lowest band. With this in mind this worked example uses bootstrap samples have been re-run for displaying the data in bands between 1 and 20 points.

This method then produces the optimal band for each integer score. Further work is required to apply these to the full distribution, as adjacent integer scores may have conflicting optimal bands. For this example the simple approach has been taken of starting at the lowest score, applying the necessary band, then optimal band for the next highest score is applied etc. This continues until the full score range has been banded. There are potentially more sophisticated and accurate methods available for applying these bandings, the simplest of which would be to carry out the same procedure but commencing at the 'centre' of the distribution.

Having performed the bootstrap sampling and the optimisation technique the following bands were obtained for this data and scorecard:

Low	High	Bandwidth
98	118	20
118	130	12
130	140	10
140	150	10
150	160	10
160	169	9
169	179	10
179	190	11
190	203	13
203	215	12
215	230	15
230	246	16

The following table shows the score distribution by these bands.

Gini: 57.81
KS: 0.42 at 160

SCORE	Count Actual	Goods Actual	Bads Actual	Bad_Rate Actual
LowScore	0	0	0	0.0%
98	162	33	129	79.6%
118	190	34	156	82.1%
130	209	60	149	71.3%
140	261	104	157	60.2%
150	235	116	119	50.6%
160	258	142	116	45.0%
169	219	126	93	42.5%
179	226	147	79	35.0%
190	238	177	61	25.6%
203	195	173	22	11.3%
215	221	199	22	10.0%
230	196	189	7	3.6%
Total	2610	1500	1110	42.5%

Measuring Confidence Levels in the Predictions of Credit Risk Models.

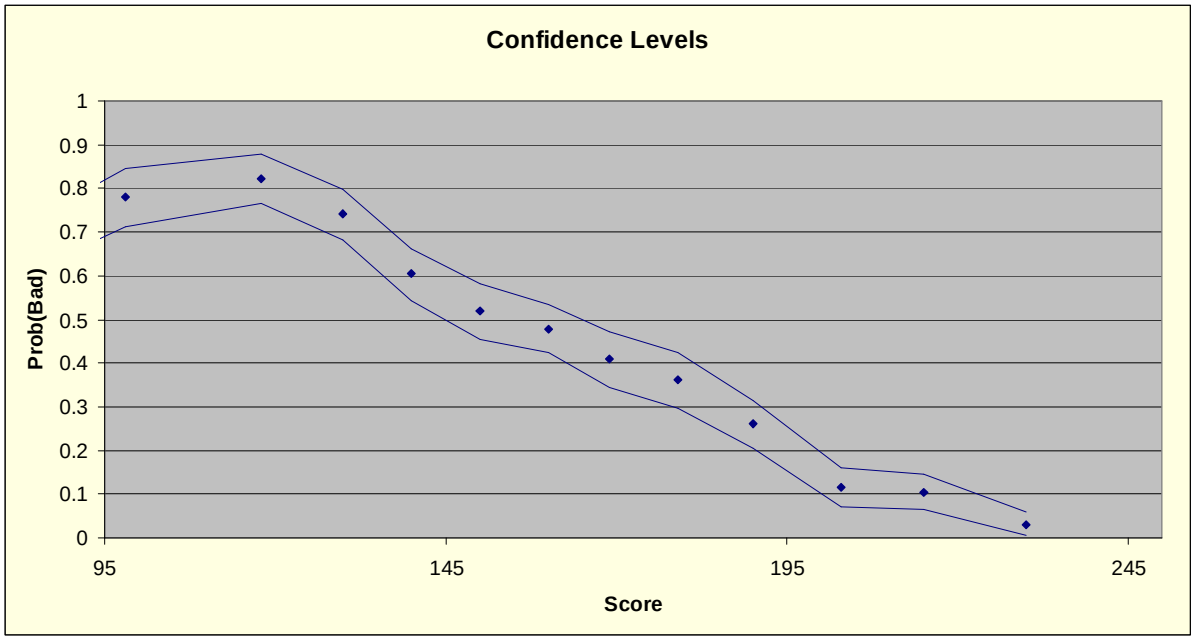
Confidence Level Calculation

Having obtained the bands in which the data should be displayed, the relatively simple task of determining the confidence can be carried out.

Having produced bootstrap samples for determining the bands we can simply re-use these samples to obtain the confidence levels required.

The confidence levels are obtained by ranking the sample bad rates in each band and then top and tailing to obtain the required confidence level. For example a 95% confidence level would require the range of bad rates in each band to be 'top and tailed' by 2.5%.

When this is done the following results are obtained:



One can see that the confidence levels around the actual bad rate points are significant. For example the records scoring between 150 and 159.99 have a bad rate of 50.6% but the 95% confident level shows that this bad rate could range from 45% to 58%.

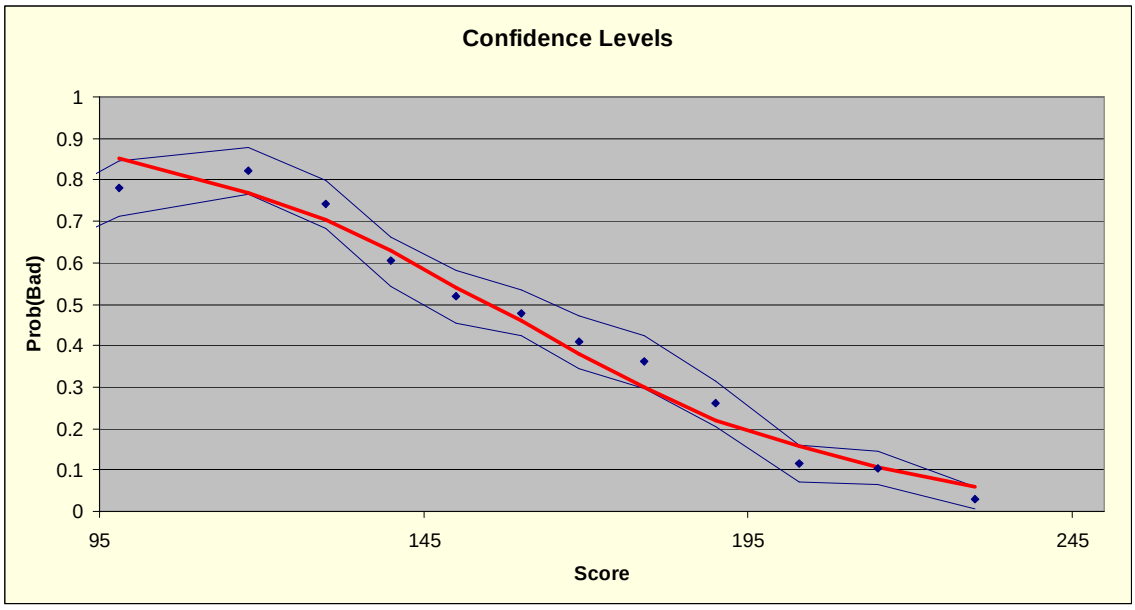
Measuring Confidence Levels in the Predictions of Credit Risk Models.

Relevance of Results

If we assume that bootstraps are producing possible distributions then the range of results shown above are the possible ranges obtained upon implementation. This could have serious impact on the way in which scorecard is used and scores are manipulated. We shall consider a few of these here.

Analysis of Predictive Accuracy

The probability of bad ranges shown in the graph above give an indication of the range of bad rates which can be obtained within a given score band. If the model is predictive then one would expected the predictive bad rate to lie within these confidence levels.



One can see from this graph that the predicted line is bound (just) within the confidence levels obtained. However areas at the extremes of the distribution may need careful consideration.

If the predicted line fell outside the confidence levels then it would suggest that the scorecard did not accurately predict in those areas.

The Placement of Cut-offs

In standard scorecard development it would be normal to produce a score distribution and bad rate curve similar to the one shown below. Despite all reservations regarding sample size issues, the score range is often presented in single point bands and then cut-offs determined by analysing the bad rate and accept rate either side of a given cut-off.

As long as population values are considered either side of the cut-off (i.e. the new accepts have a bad rate of 20%) then we can be confident in the results. However as soon as we start viewing the marginal bands either side of the cut-off we must be less sure of the actual benefits. We can see this clearly if we consider the graph above.

Measuring Confidence Levels in the Predictions of Credit Risk Models.

Imagine that the cut-off was to be drawn so that no record with a probability of bad of more than 50% was to be accepted, we can see (using the horizontal distribution of confidence levels) that the scores which incorporate a 50% bad rate span from ~145 to ~170 – somewhat less accurate than the minimal adjustments. In summary as long as population values are considered then cut-off analysis can be carried out accurately, however if the probability of bad of individual records or individual scores are being considered then range of potential values must also be brought into the analysis. In effect the movement of a cut-off by one or two points to obtain the correct bad rates in the accept population becomes at most an approximation.

Credit Limit Assignments

In a similar manner to the determination of cut-offs, the assignment of credit limits to particular scores becomes interesting in light of the confidence levels. Depending upon the size of the bands of scores being used to determine the credit limits, the confidence levels may need to be considered. This is especially true if a small number of records scoring the highest scores were given a high limit. The bad rate of accounts (in this example) scoring over 230 points could still vary between 0.6% and 6.1% - this gives a true business picture of the risk involved in awarding these high credit limits.

Calculations involving 'score'

Perhaps the area in which this analysis gives most warning is in calculations involving score. The probability of bad as predicted by the scorecard is often taken as an absolute value and then used in various calculations including, risk based pricing, other interest calculation, expected losses etc.

In these areas it is an individual record which is being assessed and as such the errors should be considered in any transformation of the score. As an example the following risk based pricing situation is analysed:

The interest rate for a product is calculated using the equation:

$$InterestRate = \frac{10\%}{P(G)^{3/2}}$$

Unless this value drops below 7%, in which case 7% is to be used.

The P(G) is devised from the score, so for example an account scoring 145 points obtains a probability of good between 0.34 and 0.46 (2dp). A score of exactly 145 would result in a P(G) = 0.37 (2dp).

This would mean that errors associated in the interest calculation would give:

Interest Rate = 7.98% (+0.22% / -0.66%)

Comparison with Standard Error

As previously stated, a drawback of the use of standard error for the estimation of confidence boundaries is its underlying assumption of normal distribution of measurements.
Previous studies have suggested this assumption to be incorrect although our results have not conclusively proved this for the data sample in use. However, we have shown that band selection is critical in the estimation of confidence levels and the band selection technique we have employed produces a non-parametric estimate of confidence limits almost as a by-product which does not rely on this assumption.

Further Work

A scorecard development sample is usually a representative sample of data extracted from a larger population. If structured sampling is performed then weights must be applied in the modelling process to reproduce the population distribution.
The process of bootstrapping becomes more complicated when using weighted records as each selected record will represent a weighted volume of cases in the resultant data set and the bootstrap sample must be selected to cater for this.

Similarly, if rejected cases are being used with the benefit of some reject inference process, then an additional weighting problem will arise.

The banding assignment mechanism presented here is simplistic and may be improved either by the suggested method of applying contiguous bands starting at some mid-point, or by applying another optimisation process.

The number of bootstraps required to produce reliable results has been questioned, but no firm answer has been suggested here.

Conclusions

We have demonstrated that accurate scorecard predictions are possible from large populations but reduced sample size inherent in the use of small score bands will tend to increase errors.

The worked example shows that some errors may be large and that errors may carry through whenever a function of score is used in a calculation.

We have demonstrated that bootstrapping techniques allow experimental analysis of errors in score (or Prob(Bad)).

We have produced a mechanism to provide an optimised set of score bands for producing confidence error limits and have demonstrated that care must be taken when these bands are not used.

Measuring Confidence Levels in the Predictions of Credit Risk Models.

This technique allows assessment of the predictive powers of a model and allows comparison between models.

The technique seems compatible (with small modifications) to other prediction methods (decision trees etc).

References

Mary L Boas. “Mathematical Methods in the Physical Sciences”. John Wiley and Sons

Corrinne Neal. “Observations on Confidence Intervals” Standard and Poor’s Singapore, April 2003