Non-Linearity of Scorecard Log-Odds

McDonald, Ross A.[1], Smith, Keith D.[1], Sturgess Matthew J. [1]& Huang, Edward X. M.

Abstract

The use of linear and log-linear (and in particular logistic regression) models for scorecard construction is near-universal. In this paper we address the question of non-linearity in the distribution of a scorecard's inferred log-odds to score relationship. Scorecards based on linear and generalised linear models are excellent and robust ranking tools, but the inferred log-odds and default probabilities are increasingly used in day-to-day business operations - within account-level strategies, for cutoff setting, in management information and for capital allocation. All are dependent upon the accurate estimation of log odds and hence probability of default, which is a quality independent of a model's ranking performance. The Gini and Kolmorogov-Smirnoff statistics that are most commonly used to evaluate scorecards reflect ranking performance only.

The nature of the variables and binning strategies frequently used within scorecards and limitations of some software packages can lead to highly skewed good / bad distributions and a non-linear log-odds relationship. The standard scorecard scaling methods assume linear transformations and can lead to significantly misleading results.

We discuss various methods for avoiding non-linearity, both at the model-build stage and in the retrospective treatment of model outputs. The discussion is illustrated with an example drawn from a scorecard generated within Lloyds Banking Group Retail Decision Science.

1. Introduction

The use of standardised scorecards for automated credit decisioning enjoys a long history in retail lending stretching back beyond the advent of computerised branch systems. As described in [Thomas et al. 2002, McNab and Wynn 2004], the widespread application of the statistical classification techniques developed by Ronald Fisher and others dates back to at least the 1950s. The arrival of credit cards in the 1960s and 70s combined with the sheer volumes of applicants and the passage of Equal Credit Opportunity legislation in the U.S. rendered such systems a necessity [Thomas et al. 2002].

To the end-user, a credit application scorecard consists of a rule set or a table of numeric weights which are summed and compared to a pre-determined acceptance threshold. More often than not these yield a three digit whole number, with low scores being interpreted as 'bad' and high scores as 'good'. Modern credit-savvy consumers are able to monitor and even influence their own scores by registering with a credit bureau and correcting inaccurate information or modifying their behaviour, though the majority are almost certainly not aware that it reflects an aggregate, transformed

[1] Retail Decision Science, Lloyds Banking Group

estimate of their expected propensity to default. This decoupling of the technical, statistical knowledge required to build a credit scorecard and the knowledge required to apply and interpret it is perhaps the single most important factor in their historical development and popularity. Traditionally, the modelling expertise required to build a scorecard could easily be outsourced to a specialist agency or bureau, with only the very largest financial institutions employing dedicated statistical modellers, often as individual specialists within larger operational teams.

The advent of large, dedicated modelling teams within retail banks is a more recent development, and has coincided with an expansion both in the type and quantity of models deployed and in the range of uses to which they are put. These are no longer confined to simple accept / reject decisions. Predicted bad rates on application scorecards are closely monitored and used to drive customer acquisition strategy to reflect business risk appetite. Behaviour scores for existing customers are used as key inputs to credit line increase and decrease strategies, in collections and recoveries, and to manage dormant accounts. A parallel breed of customer-centric value management models are used to target direct mailing campaigns to those most likely to respond, in risk-based pricing, and to route high-value customers to agents in call centres. The advent of the Basel 2 accord and the regulatory requirement to model risk and incorporate risk models in decision making has similarly led to a dramatic rise in the profile of scorecard models.

With some exceptions, the increasing sophistication with which scorecard models are deployed and used has not yet been matched by a commensurate increase in the general sophistication of the models themselves. Discriminant analysis and logistic regression, the former developed by Fisher in the 1930s [Fisher, 1936], the latter entering into widespread use in the 1980s once computing power developed sufficiently to allow fast estimation of parameters, remain the norm.

Frequently these models fail to yield a linear relationship between the predicted log-odds and the model score, a relationship implicit in the standard and widespread manner in which model weights are scaled to whole numbers. As this has no bearing on their ranking performance, they tend to remain robust and efficient ranking tools, sorting customers from high to low risk (or value, depending on the application), but may be found significantly wanting when it comes to estimating actual probability values. Nearly all the traditional metrics for evaluating scorecard performance relate to ranking performance alone, whereas the increasing sophistication of the end-user mean that decisions are often based on inferred odds ratios or propensity values. In this paper, we highlight some of the problems this may entail and their probable consequences. We also identify and test a range of possible remedies.

In Section 2 of this paper we illustrate the problem of non-linearity of model outputs against scorecard log-odds as frequently encountered in a business context. In Section 3 we briefly review the most common statistical classification methodologies used for constructing scorecards and identify the underlying causes of the problem. Section 4 explores a number of possible remedies. Finally, Section 5 summarises our findings and proposes avenues for future research.

2. The Problem

In its raw unscaled form, the output of a typical scorecard model yields the log of the odds ratio

$$\log\left(\frac{p_2}{p_1}\right)$$

Where p_1 is the probability that the event in question (e.g. credit default, fraud, take-up of a marketing offer) will occur and $p_2 = 1 - p_1$ is the probability that it will not (so a higher value for the log-odds here corresponds to a lower probability of the event's occurrence). **Figures 2.1** and **2.2** illustrate the log-odds to model-prediction relationship of a scorecard built historically within Lloyds Banking Group's Retail Decision Science division using two standard methods. For reasons of corporate confidentiality, the purpose for which the scorecard was built and score scaling are not disclosed. The plots shows the actual log-odds derived from the development data sample against the log-odds predicted by the model. The event that the model is built to predict occurs with low frequency in the sample.

The scorecard was developed in the first instance using the method of divergence maximisation via Fair Isaac's Model Builder software (**Figure 2.1**), and subsequently using a more traditional logistic regression approach implemented in SAS (**Figure 2.2**). The score distributions for the two classes are shown in **Figure 2.3** and **Figure 2.4**. The data variables were treated routinely, using weights of evidence to define a variable binning and employing a stepwise variable selection method based on incremental variable contribution. The models contain no continuous variables – all are substituted by dummy binary variables with one model coefficient for each bin. All bins for any given variable are included in the model.

In both instances, a pronounced and systematic deviation away from the line of equality for actual and predicted log-odds is observed. This deviation is approximately quadratic in shape and very significant. The curvature is observed to be more pronounced for the divergence maximisation (a possible reason for this is detailed in the next section). In both cases, the distribution for the smaller class (where the event occurs) appears approximately normal and well-behaved – that for the larger class (the event does not occur) appears irregular with heavy negative skew.

Figure 2.1: *The relationship between scaled score and log-odds for a scorecard-type model built within Lloyds Banking Group Decision Science using divergence maximisation (the plot is derived from the training sample).*

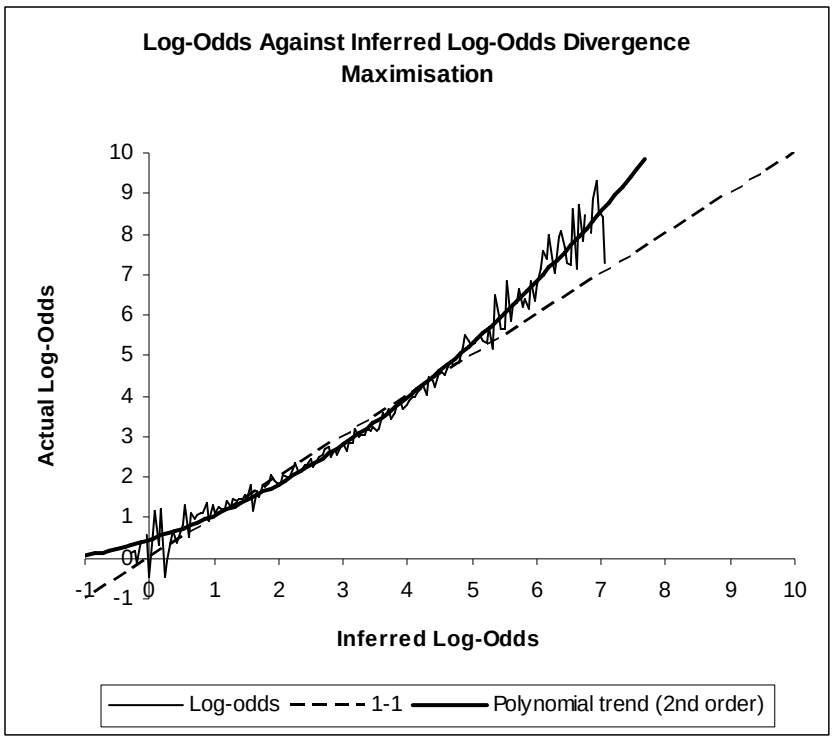


Figure 2.2: *The relationship between scaled score and log-odds for the same model built using logistic regression.*

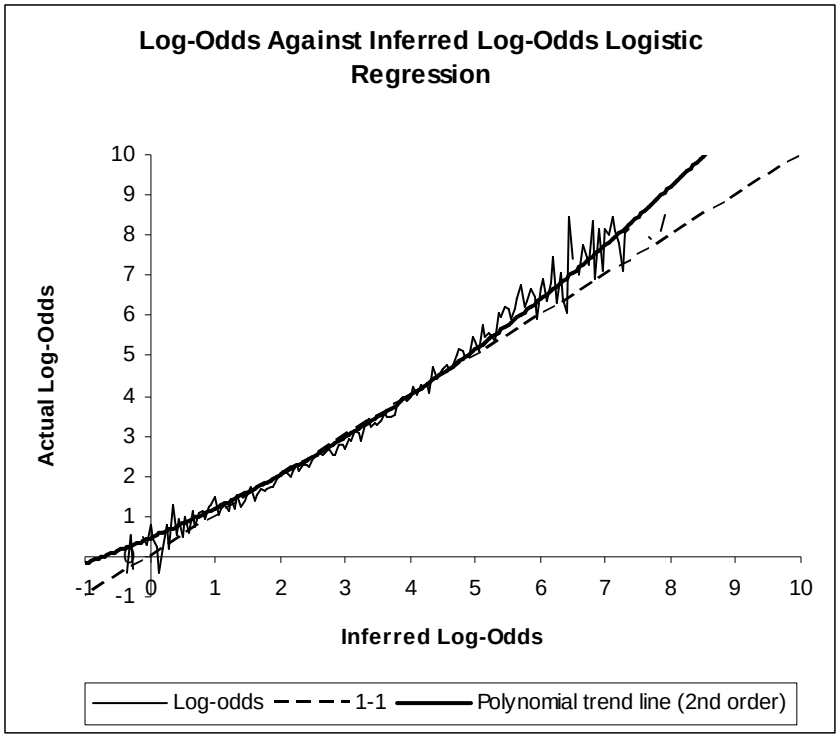


Figure 2.3: *The distribution of scores for the two predictive classes – divergence maximisation, with trendlines. Scores for the larger group exhibit a significantly right-skewed distribution.*

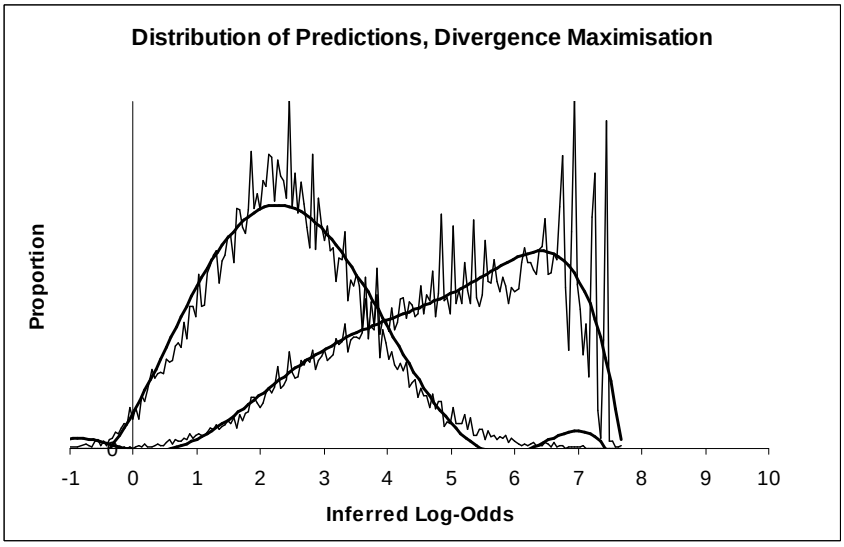
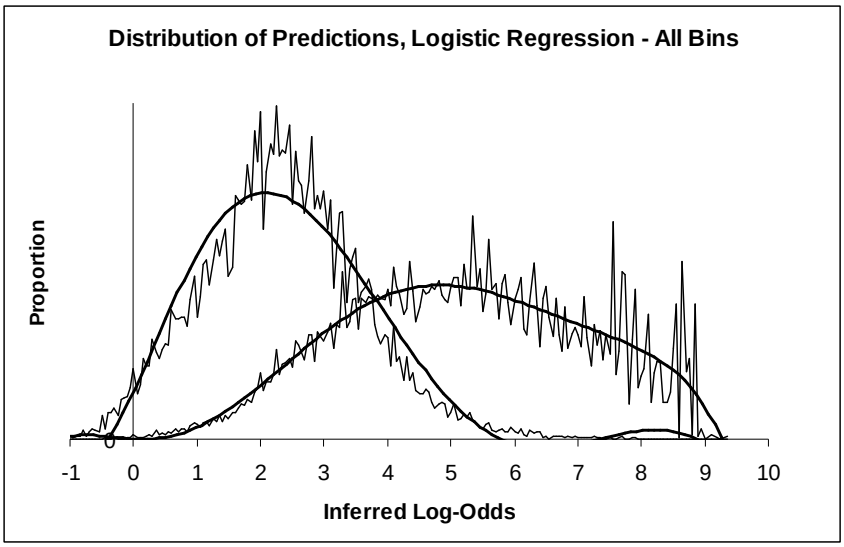


Figure 2.4: *The distribution of scores for the two predictive classes – logistic regression, with trendlines.*



In our experience, the phenomenon exhibited here occurs to a greater or lesser degree across a wide range of scorecards and data types. It is, however, natural to question whether this poses any serious problem in practice. Holdout sample ranking performances of both models measured in terms of Gini coefficient were found to be more than acceptable for the problem at hand, with that of the first model being marginally higher. Indeed any monotonic transformation of model outputs will have no impact on the ranking performance of a model and hence Gini, a fact that is exploited when output scores are routinely rescaled. If ranking performance may be considered a sufficient measure for the performance of a scorecard model where the actual predictions are not of interest, then the curvature observed here is of no practical consequence.

Historically, the end uses to which such models were put did not generally call for anything more than this (and indeed the ranking performance of a model might be considered to be more robust to changing economic or environmental conditions than its actual predictions). However, new developments in the sophistication of the way that models are used mean that the actual log-odds predictions frequently *are* of interest – for example, they may be used for cutoff setting, within account-level strategies, in management information, as a component in loss provisioning and net present value calculations, or for determining capital allocation and regulatory capital requirements across retail portfolios. The systematic inaccuracy of the log-odds curves illustrated here may therefore have very significant consequences and warrants further investigation.

3. The Underlying Causes

3.1 Standard Classification Models

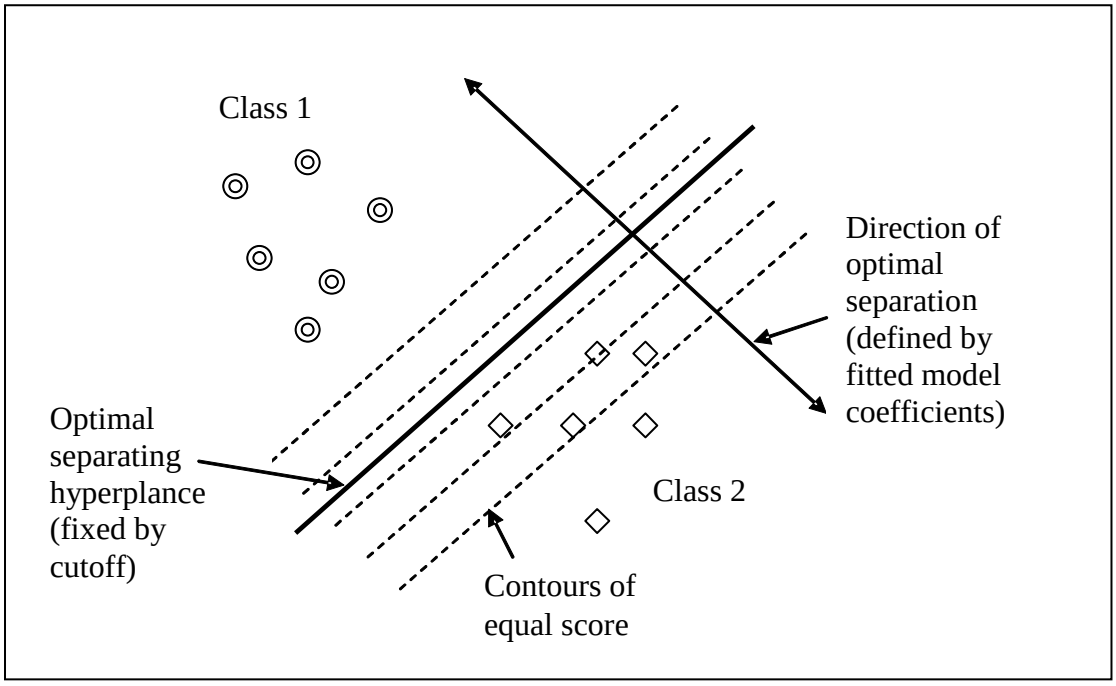
As we shall describe in this section, severe non-linearity is driven by underlying properties of the data, and aggravated by small sample sizes and correlation across separate variable bins. The models described in the previous section made use of two of the most common 2-class scorecard-building methods. These are both linear models, as the output is a linear combination of the input variables:

$$s = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Here *s* are the model scores, $\vec{\beta}$ is a vector of fitted model coefficients, $\vec{x}$ is the vector of observed measurement values and *k* is the dimensionality of the measurement space. Continuous variables in $\vec{x}$ tend to be binned and coded as 'dummy' binary variables – see e.g. [Thomas, 2009] for an explanation. This means that the range of the continuous variable is broken down into a number of exhaustive segments (say, *m*), and *m-1* binary variables are created taking the value 1 where the value of the continuous variable lies within the range and 0 otherwise (one of these variables may be dropped as it is a linear combination of the others, though this is not always done). The same substitution is often applied to categorical variables as well, with each dummy corresponding to membership of one category. If we consider only a single class, it may be assumed that each x_i is a realisation of a random variable X_i, the outcome of a set of independent 0-1 Bernouilli trials, and all of the X_i corresponding to the bins of a single binned variable are a special case of the multinomial distribution with parameter 1 (since across one binned variable, exactly one bin will take the value 1 assuming all bins are retained). The distribution of the full set of binned variables will be the sum of a set of such multinomial distributions.

As illustrated in **Figure 3.1**, the general aim of a scorecard model is to determine the direction in the multi-dimensional measurement space that optimally separates the two classes.

Figure 3.1: *A two-dimensional illustration of a predictive classification model for two classes.*



Fisher's Linear Discriminant function [Fisher, 1936] is defined as

$$\frac{(\vec{\beta}\cdot(\vec{\mu}_1-\vec{\mu}_2))^2}{\vec{\beta}'(\Sigma_1+\Sigma_2)\vec{\beta}} \qquad (3.1)$$

Where $\vec{\beta}$ is the vector of model coefficients (the intercept may be dropped without loss of generality), $\vec{\mu}_1$ and $\vec{\mu}_2$ are the means of class 1 and class 2, and Σ_1 and Σ_2 are the class covariance matrices. Maximising this function is equivalent to simultaneously maximising the separation of the means of the model scores whilst minimising their total variance – intuitively this is a sensible approach. A simple solution may be found by assuming that the data follow standard Gaussian distributions and that $\Sigma_1=\Sigma_2$, then solving directly (this is known as linear discriminant analysis). The method of divergence maximisation used to build the model in **Figure 2.1** is a more general form of this approach, dropping the assumptions of Gaussian data and equality of covariances and maximising this function directly using an optimisation algorithm (Fisher's own method amounted to constrained optimisation). **Figure 2.3** below shows the distributions of scores for the two classes for this method.

Logistic regression is an example of a generalised linear model [see McCullagh and Nelder, 1989] and proceeds from the assumption that there is an explicit linear relationship between the log odds ratio and the model covariates as follows:

$$\log\left(\frac{p_2}{p_1}\right) = \beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k$$

This formulation has the advantage of constraining *p*, which is a probability, between 0 and 1. Logistic regression models are usually fitted via maximum likelihood estimation, most commonly solved via an iterative Newton-Raphson procedure. The log-likelihood is given by:

$$\ln[l(\vec{\beta})]=\sum_{i=1}^{N}\left\{y_i\vec{\beta}^T\vec{x}_i+\log(1+e^{\vec{\beta}^T\vec{x}_i})\right\}$$

Where *i=1,…,N* indexes observations in the training sample, $y_i=0$ for class 1 and 1 for class 2, and $\vec{x}_i$ is the vector of measurements for observation *i* (including a constant intercept term). Differentiating this and setting equal to zero leads to a series of non-linear equations to solve for $\vec{\beta}$:

$$\frac{\delta\ln[l(\vec{\beta})]}{\partial\vec{\beta}}=\sum_{i=1}^{N}x_i(y_i-\frac{e^{\vec{\beta}^T\vec{x}_i}}{1+e^{\vec{\beta}^T\vec{x}_i}})=0 \qquad (3.2)$$

Recall that in scorecard construction we have typically coded all of our variables such that every member of $\vec{x}_i$ takes a value equal to either 0 or 1. In the special case where only one value of $\vec{x}_i$ can take the value 1 at any time (e.g. where our model is based on the dummies for a single binned variable), $\vec{x}_i$ for each class is a realisation of a multinomial distribution taking binary values. In this case it can be shown that a unique optimal solution to equation 3.2 may be found when each coefficient is proportional to the log odds for positive values of that variable, ie.

$$\beta_j = c \log\left(\frac{\sum_{x_i=1} y_i}{\sum_{x_i=1} (1 - y_i)} \right) \qquad (3.3)$$

Where *c* is a constant of proportionality related to the intercept. This is the Weights of Evidence [e.g. see Thomas, 2009] value for this variable, and implies that the pattern of optimal model coefficients in this case would exactly correspond to the weights of evidence pattern across variable bins. This is intuitive, as Weights of Evidence is a standard measure of the significance of each bin across a binned variable.

Where multiple variables are binned and included in a model, it is likely that some bins will be highly positively correlated. There can be many causes for this, but the most significant is probably the impact of 'missing value' correlation. This is simply because where a data is missing for a given variable it is likely to be missing across other variables for the same reason. For example, many scorecards in retail banking are based on linked-address data provided by external credit bureaux – where there is a failure to match an address for a customer, this is likely to yield missing values across a range of variables. Similarly, where a set of variables in a model relate to aspects of a customer's prior behaviour, and the majority of customers have no applicable data (e.g. they have never responded to a mailed offer, or never gone into arrears on a credit product), sets of variables that aim to capture behaviour given the event are all likely to take their default values.

Loosely speaking, the effect of bin correlation is that the coefficient weight that would previously have been applied to one bin is now distributed across several, and the pattern of coefficients for a binned variable no longer necessarily reflects the weights of evidence pattern of that variable. This can harm the interpretability of a model – in extreme cases, a raw model coefficient may even have the opposite sign to that expected, because it is being used in effect to 'dampen' the effect of an overlarge coefficient elsewhere.

This is linked to the fact that bin correlation opens up the possibility of multiple optimal solutions to the discriminant function (equation 3.1) in the case of divergence maximisation, and to the maximum likelihood function (equation 3.2) for logistic regression. This gives the model much greater leeway in finding an optimal solution, resulting in the skewed and irregular score distributions visible in **Figures 2.3** and **Figure 2.4**. The method of divergence maximisation, which seeks to maximise separation of the means while minimising total variance, appears to have shifted the mean of the larger class up the range at the expense of increasing its variance (as skewed distributions often have significantly higher variance than symmetric ones with otherwise similar dispersion). In the case of logistic regression, the irregular redistribution of coefficient weights across variable bins has led to a similar, though less pronounced skew, and an irregular distributional shape.

The impact of removing highly correlated variable bins from the logistic regression model and rebuilding is illustrated in **Figures 3.2** and **3.3** below. Here where the absolute value of the correlation coefficient of a variable pair exceeds the threshold, only the bin with the larger contribution to model performance (Gini of the single-variable model) is retained. The thresholds shown are 1, 0.5, 0.4 and 0.3. We observe that the distribution of the larger class becomes more regular as correlated bins are

progressively eliminated, and that the curvature in the log-odds relationship becomes progressively less pronounced. We also expect the removal of bins to have some impact on the performance of the model, as even highly correlated bins can add useful predictive power. The proportional impact on the model's Gini coefficient as correlation is removed is shown in **Figure 3.4**.

Figure 3.2: *The impact on the distributions of scores when bins correlated above a certain threshold are progressively removed.*

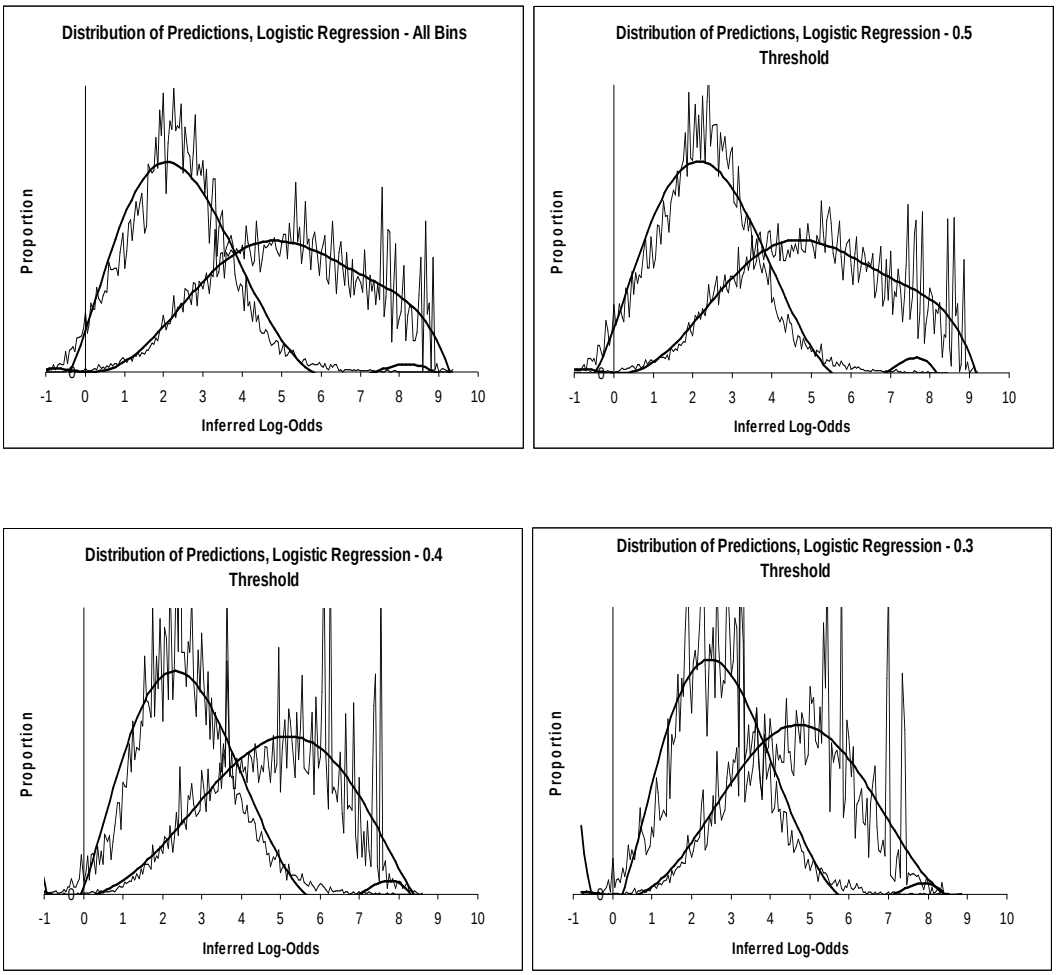


Figure 3.3: *The impact on the log-odds to score relationship when bins correlated above a certain threshold are progressively removed.*

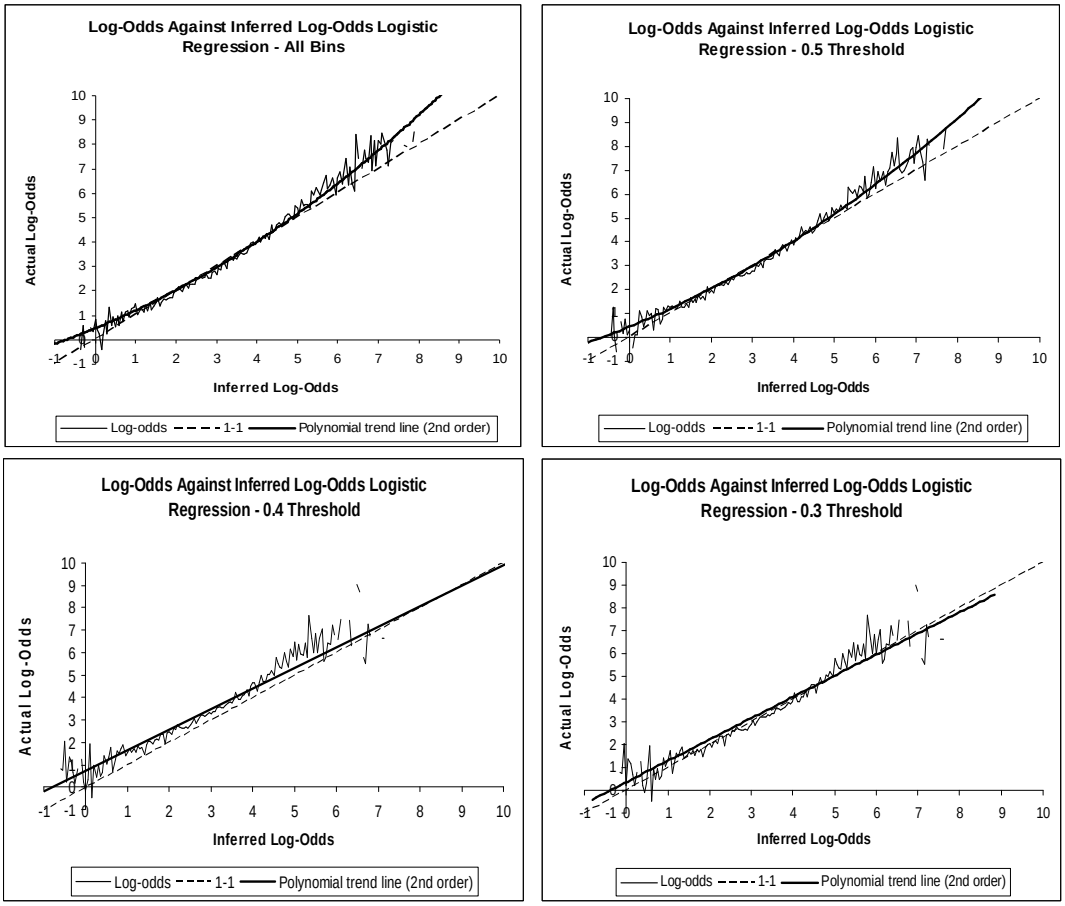
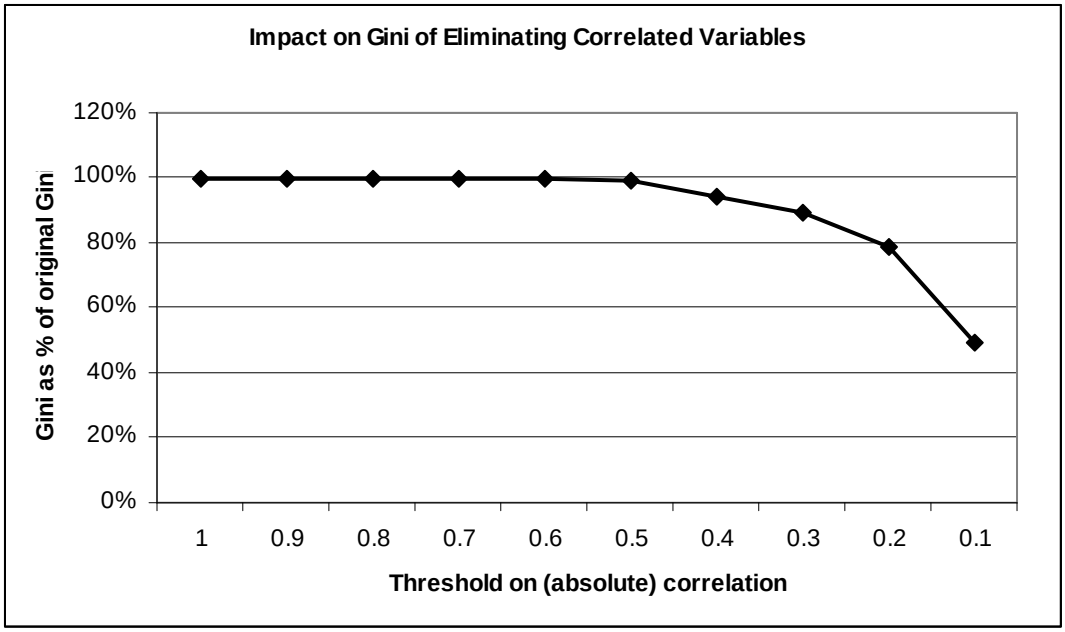


Figure 3.4: *The impact on the Gini coefficient of the logistic regression model of progressively removing correlated bins (Gini shown as a percentage of the Gini of the model that includes all bins). Gini falls by approximately 50% once all correlations above 0.1 are removed.*



3.2 Quadratic non-linearity

The strong curvature observed in the log-odds to score relationship of our models is related to the variances of the score distributions. The following argument illustrates how this relationship may work under the assumption that the scores for the two classes are normally distributed with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2. Then

$$\frac{P(s \mid Class = 2)}{P(s \mid Class = 1)} = \frac{\sigma_1}{\sigma_2} \exp\left[\frac{1}{2} \left(\left(\frac{s-\mu_1}{\sigma_1} \right)^2 - \left(\frac{s-\mu_2}{\sigma_2} \right)^2 \right) \right]$$

Therefore the log-odds relationship for a given score is:

$$\log\left(\frac{P(Class = 2 \mid s)}{P(Class = 1 \mid s)} \right) = \log\left(\frac{P(Class = 2)}{P(Class = 1)} \right) + \log \frac{\sigma_1}{\sigma_2} + \frac{1}{2} \left(\left(\frac{s-\mu_1}{\sigma_1} \right)^2 - \left(\frac{s-\mu_2}{\sigma_2} \right)^2 \right)$$

$$= \log\left(\frac{P(Class = 2)}{P(Class = 1)} \right) + \log \frac{\sigma_1}{\sigma_2} + \frac{1}{2} \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} \right) + \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2} \right) s + \frac{1}{2} \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) s^2$$

$$= as^2 + bs + c$$

Where

$$a = \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right),$$

$$b = \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right),$$

$$c = \log\left(\frac{P(Class = 2)}{P(Class = 1)}\right) + \log\frac{\sigma_1}{\sigma_2} + \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right)$$

Although this argument strictly applies only where scores are approximately normally distributed, and we have already observed that correlation in the data can lead to highly irregular score distributions, it does nonetheless help to illustrate the mechanism through which the curvature arises. The degree of curvature is here a function of the coefficient *a*, which is determined solely by the difference between the reciprocal variances of the score distributions. Because *a* is related to the reciprocal variances, larger values of *a* are likely to arise when the variances are smaller. As small score variances are likely to be a feature of scorecards that discriminate well, this leads us to the counterintuitive finding that stronger scorecards may be more prone to curvature than weaker ones.

Although the progressive elimination of correlated bins appears to lead to progressively more regular score distributions, it is clear from **Figure 3.2** that correlation alone cannot fully explain the difference in their variances. Recall that we posited that a single variable bin could be considered the outcome of a series of independent Bernouilli trials such that all the bins corresponding to a binned variable are multinomially distributed. Strictly speaking, this is only the case when we consider the two classes separately, since assuming that the variable has any discriminatory power the parameter $p_i = prob(x_i=1)$ is conditional on *y*, the class that we are considering. It can be shown that if we consider the total score as the weighted sum of the components of a series of multinomially distributed variables, the variance of this distribution reduces to the weighted sum of a sequence of significant components of the form $p_i(1-p_j)$.

When $i \neq j$, these components are directly driven by correlation between separate variables, which again illustrates how correlation can contribute to differences between the score variances. When $i=j$, we are considering a single variable bin. If the variable had no discriminatory power whatsoever, then p_i would have no dependence on *y* and the contribution to the score variances of this variable would be equal. Conversely, if the bin discriminated perfectly, p_i would take the value 0 for one class and 1 for the other, again leading to equal score variance. Finally, in the very unlikely case that a variable bin has partial discriminatory power, but the proportions of 1s and 0s are equal and match the overall class proportions which are also equal, then the contributions to the score variances will be equal.

The above conditions are very unlikely to be met for a significant modelling variable in practice, as a binned variable in a model is most likely to have a least some discriminatory power with the class proportions likely to be somewhat (and more than likely very) unequal. Note that small sample sizes for a particular class within a bin may also aggravate the differences, due to the increased likelihood of error in the estimation of p_i.

To summarise, some difference in the variances of the score distributions for the two classes is almost inevitable in any practical modelling situation. The presence of bin correlation will further magnify the effect due to the presence of the cross-variable terms in the variance equation and the propensity to skew the score distributions. And the difference in score variances is the most likely cause of the curvature in the log-odds to score relationship.

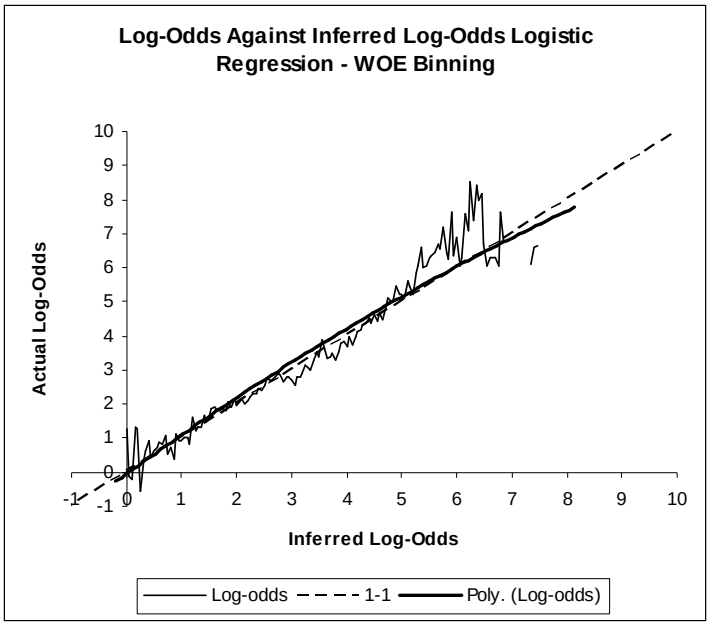
4. Remedies

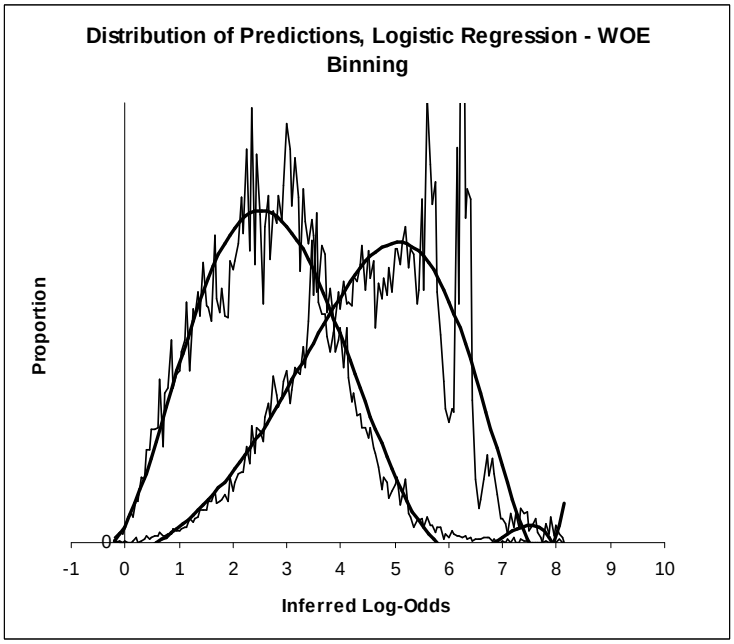
The arguments above relating to the causes of the curvature in the log-odds to score relationship suggest several causes and therefore a number of remedies. We have already shown that eliminating individual correlated bins from a model can lead to more regular score distributions, reducing skewness and the differences in variance, but not all software packages commonly used for scorecard construction will allow individual bins to be dropped. As evident in **Figure 3.4**, however, the removal of correlated bins does compromise the performance of the model, as even bins that are highly correlated may provide new information.

Another solution may lie in the modelling technique employed. For example, as linear discriminant analysis proceeds from an assumption of equal covariances in equation 3.1, and covariance terms contribute to the non-linearity, it is possible that a scorecard built in this way would prove more robust to the curvature than one based on divergence maximisation, which drops this assumption (though it is equally likely that the latter model would perform better in terms of standard measures based on ranking performance).

Other solutions may lie in the methods used for binning. The problem of estimation error within bins may be improved by ensuring that bins do not become too small. An alternative approach to binning uses the weights of evidence within bins directly in the model. Rather than replace each variable with a set of dummy variables denoting membership of each bin, the values weights of evidence values corresponding to each bin are substituted for each actual value in the corresponding variable. The resulting continuous variables are then fed into the model, in effect enforcing the condition in equation 3.3 that the model covariates will exactly reflect the weights of evidence pattern. **Figure 4.1** below illustrates the resulting impact on the data we are considering here using the logistic regression model.

Figure 4.1: *The log-odds to score relationship and score distributions when weights of evidence are enforced in the logistic regression model.*





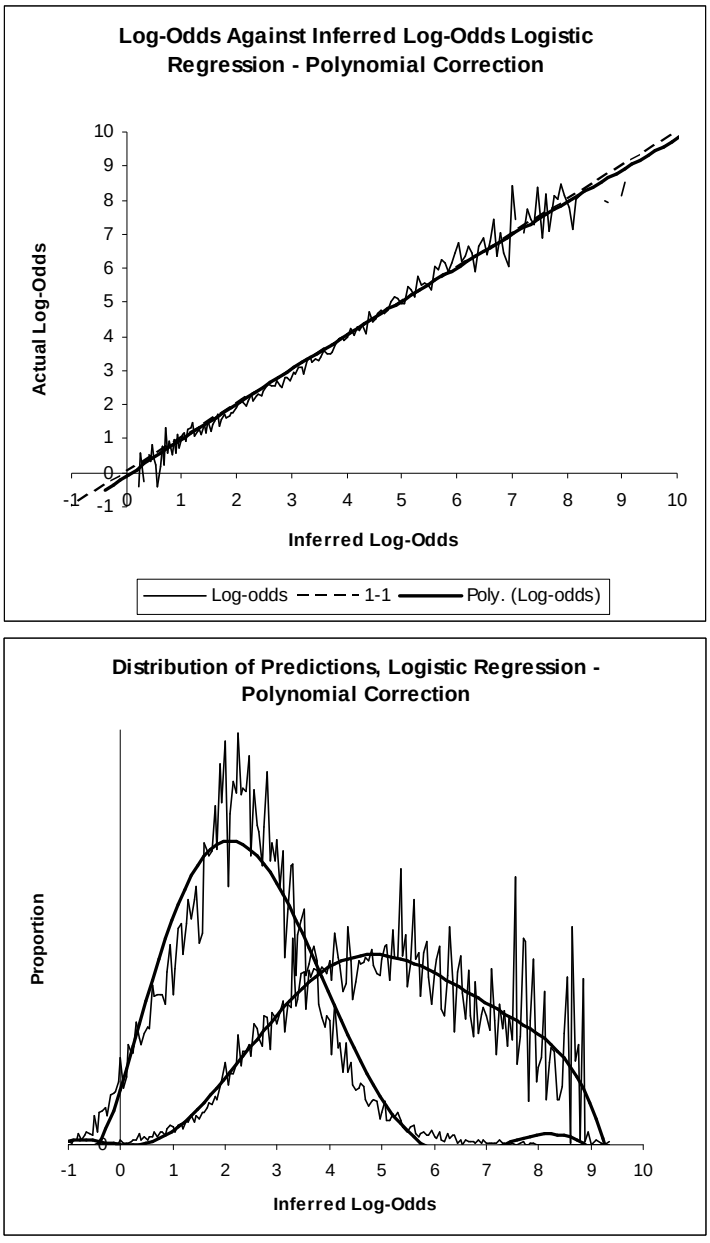
The score distributions are more regular under this method despite some polarisation among higher score values, though they remain somewhat skewed, possibly because correlation at a variable level (as opposed to at a bin level) is still possible. The curvature of the log-odds is significantly improved. However the Gini is weaker for this model – only 90% of that for the fully binned model, a reduction in performance that is unlikely to be acceptable in a business environment.

Our preferred corrective measure involves applying a retrospective non-linear transformation to the model scores to correct the curvature. Recall that any monotonic (not only linear) transformation of score values can be applied without having any impact on the model's ranking performance (and hence the Gini coefficient) at all. A polynomial correction is not universally monotonic, but is likely to be very close to being so across the range of interest. **Figure 4.2** below illustrates the impact on the logistic regression model of applying such a correction – the score distributions

remain irregular, but the curvature is corrected, and there is only a minimal impact (<1% reduction) on Gini which may be attributed to rounding error.

The only drawback of this method in a business environment is operational, because the correction must be applied after the scores are summed. In many situations this will not be an issue, but some of the decisioning systems on which scorecard models are implemented will not allow retrospective non-linear score transformations of this kind after an additive score has been calculated, and even where this is possible a significant number of end users of the score may need to be re-educated in its use.

Figure 4.2: *The effect of applying a retrospective polynomial transformation to the logistic regression model to align the scores.*



5. Conclusions And Future Work

In this paper we have described and illustrated the common problem of a systematic mismatch between a scorecard model's predictions and the actual log-odds in the data to which it is applied. It is our hope that this discussion will lead to a greater understanding and awareness of the issue. We identified the operational risk that this can pose in retail banking given the wide range of uses to which the actual inferred predictions of scorecard models are put.

We have identified a number of underlying causes of the problem – these include small data issues, the underlying properties of the data, and in particular correlation across variable bins leading to skewness of the score distributions. Although it is our hope that the arguments above will help to elucidate the problem, our discussion is by no means intended to be fully mathematically rigorous or exhaustive. In particular, we have only considered the case of scorecards where all variables are binned, and have what happens when continuous variables are included in a model. Our investigations would suggest that while continuous variables can add to the ranking performance of a model, they can have a severe and unpredictable effect on the log-odds to score relationship (unless tightly constrained, as with the weights of evidence binning method described earlier). Our discussions are also limited to linear and generalised linear models used to distinguish between two classes, though this covers the majority of models of this type used in retail banking.

Among the remedies that we investigated, we believe that applying a retrospective non-linear transformation to the model scores holds the most promise, though this may pose operational issues in some cases. Other remedies are possible, but many will have some detrimental impact on a model's performance due to the existence of a trade-off between discriminatory power and prediction accuracy.

Acknowledgements

The authors wish to acknowledge the assistance provided by Bob Samra of Lloyds Retail Decision Science in explaining the non-linearity effect, and of Richard Norgate and Rae Miller, also of Lloyds Retail Decision Science, for supporting this research.

References

Fisher, R. A. [1936] The Use of Multiple Measurements in Taxonomic Problems *Annals of Eugenics 7 179-188*

Hand, D. J. [2002] Good Practice in Retail Credit Scorecard Assessment *J. Oper. Res. Soc 56 1109-1117*

McCullagh, P. and Nelder, J. A. [1989] Generalized Linear Models CRC Press

McNab, H. and Wynn, A. [2004] Principles and Practice of Consumer Credit Risk Management, CIB Publishing, Canterbury

Thomas, L. C. [2009] Consumer Credit Models: Pricing, Profit and Portfolios OUP, Oxford

Thomas, L. C., Edelman, D. B. and Crook, J. N. [2002] Credit Scoring and its Applications SIAM, Philadelphia