Signal Detection for Credit Scoring Practitioners

Ross Gayler
Equigen Consulting

The ROC curve is useful for assessing the predictive power of risk models and is relatively well known for this purpose in the credit scoring community. The ROC curve is a component of the Theory of Signal Detection (TSD), a theory which has pervasive links to many issues in model building. However, these conceptual links and their associated insights and techniques are less well known than they deserve to be among credit scoring practitioners.

The purpose of this paper is to alert credit risk modelers to the relationships between TSD and common scorecard development concepts and to provide a toolbox of simple techniques and interpretations.

History of the Theory of Signal Detection

- Measure of ability to detect a signal in noise
- Statistical decision theory
- Communications engineering
- Psychology
- Other applications: X-ray interpretation, military monitoring, industrial monitoring, information retrieval
- Credit scoring

The Theory of Signal Detection is about quantification of the ability to detect a signal embedded in noise.

Signal and noise may be defined very generally. The treatment of "signal" in TSD is equivalent to requiring a binary outcome. The treatment of "noise" is equivalent to saying that sometimes a "no signal" condition will be indistinguishable from a "signal" condition.

The original basis came from statistical decision theory (e.g. Wald, 1950).

TSD was developed in the context of communications engineering (Peterson, Birdsall, & Fox, 1954) for quantifying the ability of a receiver to detect a signal, such as a pure tone, in white noise. This was rapidly imported into perceptual psychology (Tanner & Swets, 1954) for application to similar tasks. From there it moved into cognitive psychology, quantifying the strength of memory traces (Norman & Wickelgren, 1965). In this progression across domains the signal and noise constructs became progressively less observable and more abstract.

TSD has been applied to other areas where quantification of sensitivity or discriminability is important. Swets and Pickett (1982) are primarily concerned with interpretation of radiological images, but also provide an extensive bibliography of applications in military monitoring, industrial monitoring, and information retrieval.

Some elements of TSD (e.g. the ROC curve) are known in credit scoring but the broader range of linkages is not so widely known among model builders.

Detection & Response

- TSD separates detection and response
- Detection is modelled as a mapping of entities onto a decision axis
- Detection is inherently probabilistic because of the effect of noise on the mapping
- Response is modelled by a criterion on the decision axis
- This looks like credit scoring

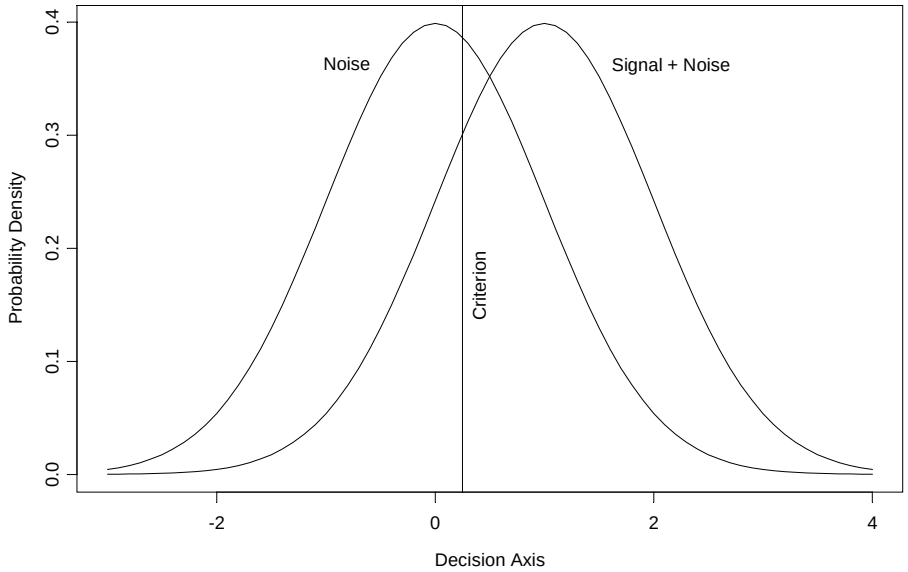
The motivation for TSD is to characterise the accuracy of an empirical process that attempts to indicate whether a signal is present. TSD models this empirical process with separate components for detection and response.

Detection maps each case onto a decision axis. The presence of noise maps each fixed signal level to a distribution of values on the decision axis. (Interpret that as metaphorically as you wish.) This makes the detection process inherently probabilistic. That is, sometimes a "signal" case looks like a "no signal" case and vice versa.

The response process is modelled by a criterion level on the decision axis. Any case above the criterion causes a "signal" response.

This looks like credit scoring. The signal is "creditworthiness". The scorecard maps cases onto the decision axis (the score). The criterion is the cutoff and the responses are labeled "Accept" and "Reject".

Mapping & Criterion



An example of the mapping and criterion. To recast it in credit scoring terms:

The decision axis is the score.

The criterion is the cutoff.

The Noise distribution is the distribution of scores of known Bad cases.

The Signal+Noise distribution is the distribution of scores of known Good cases.

Sensitivity & Response Bias

- Optimal response criterion depends on the payoff matrix
- Optimal response policy may be to ignore the detection component
- Sensitivity quantifies how well the mapping separates the distributions
- TSD allows separate quantification of sensitivity & response bias
- Other accuracy measures confound the two

Given a fixed detection mapping, the optimal criterion depends on the 2×2 (actual vs. response) payoff matrix and the signal probability.

For some payoff matrices and signal probabilities the optimal criterion is extreme, effectively ignoring the information from the case. For example, if the payoff is 100 for responding "Signal" when the signal is actually present and 0 for the other three cells, the optimal policy is to always respond "Signal".

TSD allows separate quantification of the sensitivity and response bias. Sensitivity quantifies how well the mapping separates the distributions. Other accuracy measures confound the two components. For example, percent correctly classified and average bad debt loss are functions of the sensitivity, response bias, and portfolio odds.

In credit scoring, a bottom-line figure is arguably the only one that really matters. However, reliance on bad debt (or an equivalent) as a performance measure runs the risk of making the conclusions specific to the economics of the portfolio and one decision strategy, making comparisons across portfolios and strategies difficult.

Sensitivity: *d'*

- Ignore response bias in credit scoring because it is completely controlled
- *d'* is the TSD sensitivity measure
- Separation of the distributions along the decision axis (in standard deviations)
- Assumes normality and equal variance
- Interpret *d'* as a statistical effect size
- Normality not an onerous assumption

In psychology the decision axis and response bias are not observable. Only the cases and responses are known. The emphasis in credit scoring is different because the score is directly observable and the response criterion is exactly controllable. Therefore, we can ignore response bias and concentrate on sensitivity for what it tells us about the data and scorecard.

The principal TSD measure of sensitivity is *d'*. This is defined as the separation of the distributions on the decision axis (in units of standard deviations). This definition follows from a theoretical assumption that the distributions are normal and of equal variance.

Interpreting *d'* as a statistical effect size measured in standard deviations provides extra grip for intuition. A typical Australian application scorecard might have *d'* = 1 (not a remarkably high separation when interpreted as a hypothesis test). Wilkie (1992) provides a comparison of sensitivity measures commonly used in credit scoring.

The assumptions of normality and equal variance are not particularly onerous because they are not critical and where they are manifestly broken the departure from the assumptions provides interesting information.

Observable Data

- TSD assumes decision axis unobserved
- *d'* calculated from observable data

  Hit rate = $H = P(response=Y \mid case=Y)$

  False Alarm rate = $F = P(response=Y \mid case=N)$

		Response	
		Y	N
Case	Y	Hit	Miss
	N	False Alarm	Correct Rejection

In psychology the decision axis and response bias are not directly observable. Only the responses and identities of the cases and are observable. Therefore, the sensitivity has to be calculated from the observable data.

The observable data for a fixed signal and response criterion consists of a 2×2 table with the cells conventionally labeled as: Hit, Miss, False Alarm, and Correct Rejection.

The number of cases of each type is known (being controlled stimuli in psychology and known outcomes in credit scoring). Therefore the table can be summarised by two quantities; conventionally the Hit rate ($P(response=Y \mid case=Y)$) and the False Alarm rate ($P(response=Y \mid case=N)$).

Calculation of *d'*

- Sensitivity is a function of H and F
  - Higher H implies higher sensitivity
  - Lower F implies higher sensitivity
- sensitivity = $v(u(H) - u(F))$
  - $v()$ and $u()$ are monotonic functions
- $d' = z(H) - z(F)$
- *d'* invariant under monotonic transformations of the decision axis

A reasonable measure of sensitivity should be calculated as a function of the Hit and False Alarm rates. For a given False Alarm rate a higher Hit rate implies higher sensitivity. For a given Hit rate a lower False Alarm rate implies higher sensitivity.

Note that the Hit and False Alarm rates (and hence the sensitivity) are independent of the portfolio odds. This is equivalent to saying that the sensitivity is a function of the interaction of the signal and detection system occurring at the level of the single case.

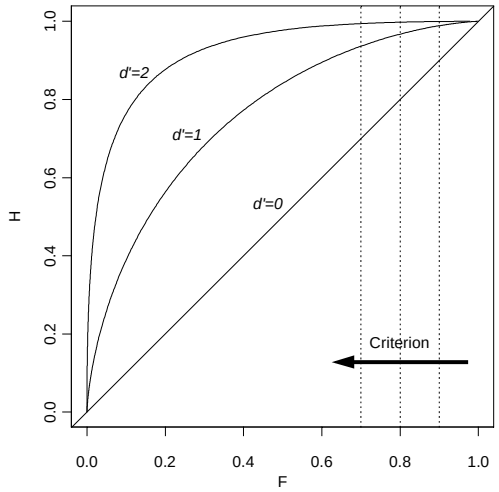
Macmillan and Creelman (1991, p. 13) argue that sensitivity measures should be of the form $v(u(H) - u(F))$ where $v()$ and $u()$ are monotonic functions.

Given the assumed distributional model, the criterion corresponding to the rates *H* and *F* can be identified with points on the respective cumulative distributions. The function $u()$ locates the criterion on the decision axis relative to each of the distributions. Thus the term $u(H) - u(F)$ can be interpreted as the distance between the two distributions along the decision axis. The function $v()$ is used to give desired scaling properties to the distance.

$d' = z(H) - z(F)$ where $z()$ is the inverse cumulative normal distribution function. This calculation is invariant under monotone transformations of the decision axis because it is based on the rates rather than the scores. Contrast this with the measure *D* (Wilkie, 1992, p. 126) which relies on calculating the standard deviation of the scores. The exact distributions for TSD are irrelevant to the extent that they can be made normal by transforming the decision axis.

ROC Curve

- Assessment of sensitivity at multiple criteria



d' is calculated at a single criterion setting. The ROC (Receiver Operating Characteristic) curve allows assessment of sensitivity at multiple criterion settings. The ROC plots *H* against *F* as the criterion is swept along the decision axis. *H* and *F* increase as the criterion value decreases (the cutoff is lowered).

The more sensitive the detector the further the curve is from the diagonal. A curve on the diagonal indicates zero sensitivity (the two distributions on the decision axis are coincident). Curves below the diagonal (corresponding to negative values of *d'*) indicate that the response is reversed.

When the ROC curve is plotted on probability axes it is difficult to estimate the sensitivity by eye. Comparison between curves consists of determining which is further from the diagonal.

The example shows isosensitivity curves for two detectors (scorecards) of different sensitivity. The curves trace out points of constant sensitivity because the detectors are simulated with data from normal distributions of equal variance.

If the assumptions do not hold the curves need not represent constant sensitivity and it is possible for them to cross. In this case the identity of the more sensitive detector varies as a function of the criterion.

The shape of the ROC curve is important when the implemented criterion may vary over a range or be extreme (e.g. in fraud models).

Area Sensitivity Measures

- Area under curve (*P(A)*) is global sensitivity
- Summary measures hide interesting features
- $d' \approx 2^{0.5} \times z(P(A))$
- % correct in two-alternative forced choice
- Gini is a linear transform of P(A)
- P(A) related to other statistics based on concordant pairs : Mann-Whitney U, c, Somers' D, Goodman-Kruskal Gamma, Kendall's Tau-a

The area under the ROC curve summarises the sensitivity of the detector over the range of criteria. A larger area implies a curve further from the diagonal and higher sensitivity. Summary figures may hide interesting features!

P(A) is the area under the curve (which must be between zero and one).

$d' \approx 2^{0.5} \times z(P(A))$ (Swets & Pickett, 1982, p. 33)

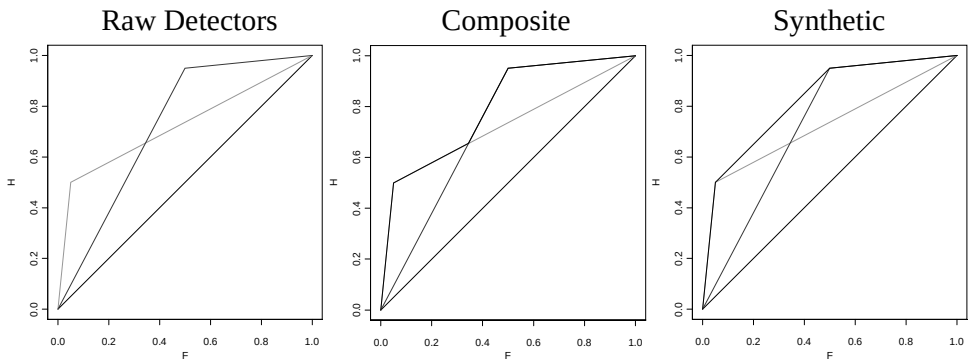
P(A) can be interpreted as the proportion correct in a two-alternative forced choice (Macmillan & Creelman, 1991, p. 125). That is, choose a known Good at random and choose a known Bad at random, then P(A) equals the probability that the order of scores is concordant with the known outcomes. Note that this is independent of the portfolio odds.

The Gini coefficient (*G*) is the area between the curve and the diagonal as a fraction of the area above the diagonal ($P(A) = G/2 + 0.5$) (Hand, 1997, p. 134).

P(A) is also related to other statistical measures of association/correlation that are based on the proportion of concordant pairs These relations may be useful as aids to intuition or as computational shortcuts. For example, SAS PROC LOGISTIC produces classification summary statistics including *c* (SAS Institute, 1989, p. 1091) which is equivalent to *P(A)*. Thus, a scorecard developer using logistic regression can easily obtain *P(A)*, and thus the Gini coefficient or *d'*, from the standard SAS output.

Composite & Synthetic ROC

- Maximise *P(A)* by combining detectors
  - Composite: deterministically switch detectors
  - Synthetic: probabilistically choose a classifier



Given multiple detectors (scorecards) their ROC curves may cross. The identity of the most sensitive detector depends on the criterion value.

A detector system with greater total sensitivity than the individual detectors can be constructed by combining individual detectors.

Composite: The composite ROC curve is the concatenation of the dominating segments of the individual ROC curves. The detector to use is a deterministic function of the desired False Alarm or Hit rate. Choose the detector with the higher sensitivity at the desired criterion. This is conceptually similar to the standard notion of having separate scorecards for subpopulations except that the implementation partition is chosen after the scorecards are developed and need not be identical to the scorecard development partition.

Synthetic (Scott, Niranjan, Melvin, & Prager, 1998): Each point on an individual ROC curve represents a classifier (a detector with a fixed criterion). The synthetic ROC curve is the convex hull of all the classifiers implied by the individual ROC curves. The classifier to use is a probabilistic function of the desired False Alarm or Hit rate. Calculate the vertices of the convex hull. Locate the desired criterion on the segment of the convex hull joining the two adjacent vertices. Randomly choose one of the classifiers corresponding to the vertices with probability proportional to the position of the criterion on the segment relative to the vertices. Apply the randomly chosen classifier to the case. (Note that as this is based on classifiers with fixed criteria the detectors need not be scorecards or other models able to use a variable criterion.)

Transformed ROC Curve

- Easier to interpret on *z()* transformed axes
  - Slope is ratio of standard deviations
  - Intercept is *d'* (if slope = 1)
- Plot transformed ROC minus the diagonal
  - Plot 2×2 *d'* as a function of criterion
- Slope = 1 equivalent to constant sensitivity
- Scores tend to have constant sensitivity
- Predictors less likely to have constant *d'*

The ROC curve on probability axes is curved, making estimation by eye difficult. If the assumptions of equal variance normal distributions are met the ROC curve will be a straight line on *z()* transformed axes. The curve is easier to interpret when transformed to a straight line.

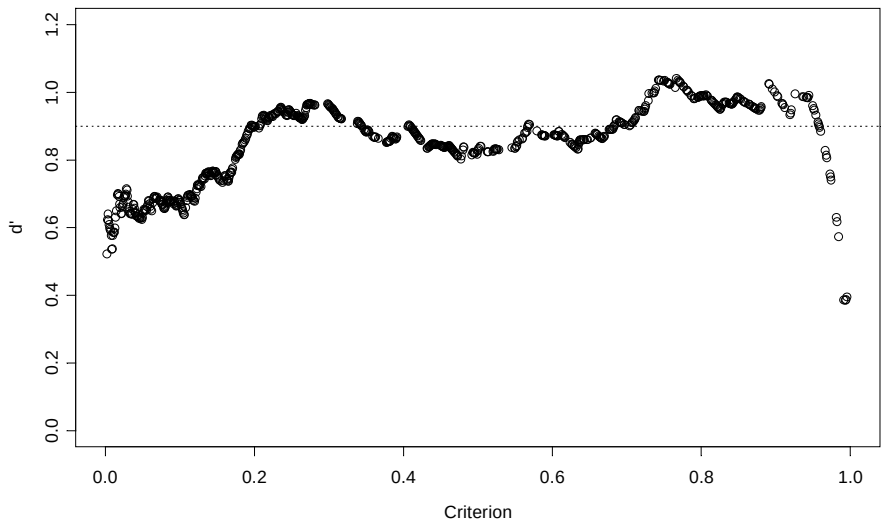
The slope of the line equals the ratio of the standard deviations of the distributions. If the slope = 1 the intercept is *d'*. If the slope is not 1 then having a single sensitivity value loses some meaning. The line could even cross the diagonal, in which case the responses are reversed for the segment below the diagonal.

The major diagonal indicates zero sensitivity. This can be taken as a null model of the detection process. Plotting the actual ROC curve minus the diagonal is equivalent to plotting the actual sensitivity as a residual with respect to the null model. This is equivalent to plotting *d'* calculated from the 2×2 table as a function of the criterion value (i.e. sensitivity as a function of cutoff).

Application and behaviour scores tend to give approximately flat lines when plotted this way. That is, for many scorecards the sensitivity is roughly independent of the placement of the criterion.

This plot can be applied to any continuous predictor, so it can be used with individual predictors. These more often deviate from flat, straight lines. I tend to use use this as a diagnostic plot for screening predictors.

Example *d'* Curve



The curve shows *d'* calculated at every possible value of the criterion.

The curve is relatively flat over most of the range and centred around *d' = 0.9* (which is reasonable for an application scorecard).

The extreme values at the top of the criterion range are typical because this is an unsmoothed, empirical plot.

The slope at the bottom end warrants investigation. An examination of the plot of outcome odds by score shows that it is flat below 0.2

Relationship of *d'* to *d**

- TSD not sensitive to choice of distribution
- Use logistic instead of normal
- Decision axis is log odds
- Score × log odds is approximately linear
- Sensitivity (*d**) is log odds ratio
- Cumulative logistic & normal distributions are approximately linearly related
- $d^* = log_2(odds\ ratio) \approx 2.55\ d'$

TSD analyses are not particularly sensitive to the true shapes of the distributions. Almost any unimodal distribution on the decision axis will produce similar-looking results. This gives us the freedom to examine other distributions.

The logistic distribution illuminates some interesting relationships to other areas of statistics. If logistic distributions are assumed then the decision axis is measured in log odds (Macmillan & Creelman, 1991, p.24).

The relationship between score and outcome log odds is typically approximately linear. Therefore, the logistic assumption is approximately equivalent to using linearly rescaled scores.

The sensitivity measure is $d^* = logit(H) - logit(F) = log\ odds\ ratio$. That is *d** is the log of the ratio of the outcome odds of all cases above the criterion to the outcome odds of all cases below the criterion.

The cumulative logistic and normal distributions are approximately linearly related, except at extreme probabilities (Cox & Snell, 1989, p. 21). This leads to the approximate relationship: $d^* = log_2(odds\ ratio) \approx 2.55\ d'$

(I use base 2 logs for convenience.)

Reject Inference Heuristic

- Flat *d** curve gives constant ratio of accept odds to reject odds across cutoff values
- Reject inference heuristic: typical ratio of accept odds to reject odds 3:1 ~ 4:1
- Heuristic consistent with typical *d**
- Ratio should depend on the sensitivity of the previous application detector
- Decision reduces and slopes *d** of accepts

A constant *d** value across cutoff values means a constant ratio of odds of accepts to odds of rejects. (Many scorecard builders find the expected constancy of this ratio surprising.)

Scorecard builders have a heuristic that the ratio of the odds of the accepts to the inferred odds of the rejects is typically in the range 3:1 to 4:1. This heuristic would follow from the typical range of *d** values and the constancy of *d** across cutoff values.

This explanation for the heuristic shows that the ratio should depend on the sensitivity of the previously used detection system. A coin toss would result in a 1:1 ratio of accept odds to reject odds. A very sensitive detector would result in a higher than typical ratio.

Unfortunately, the *d** seen in the accepts cannot be used directly to estimate the odds of the rejects. The *d** of the accepts is lower than for the population because of the reduced range of odds following from removal of the rejects. Also the *d** curve should have a positive slope because differential truncation of the good and bad distributions induces unequal variances (assuming that they are equal in the population).

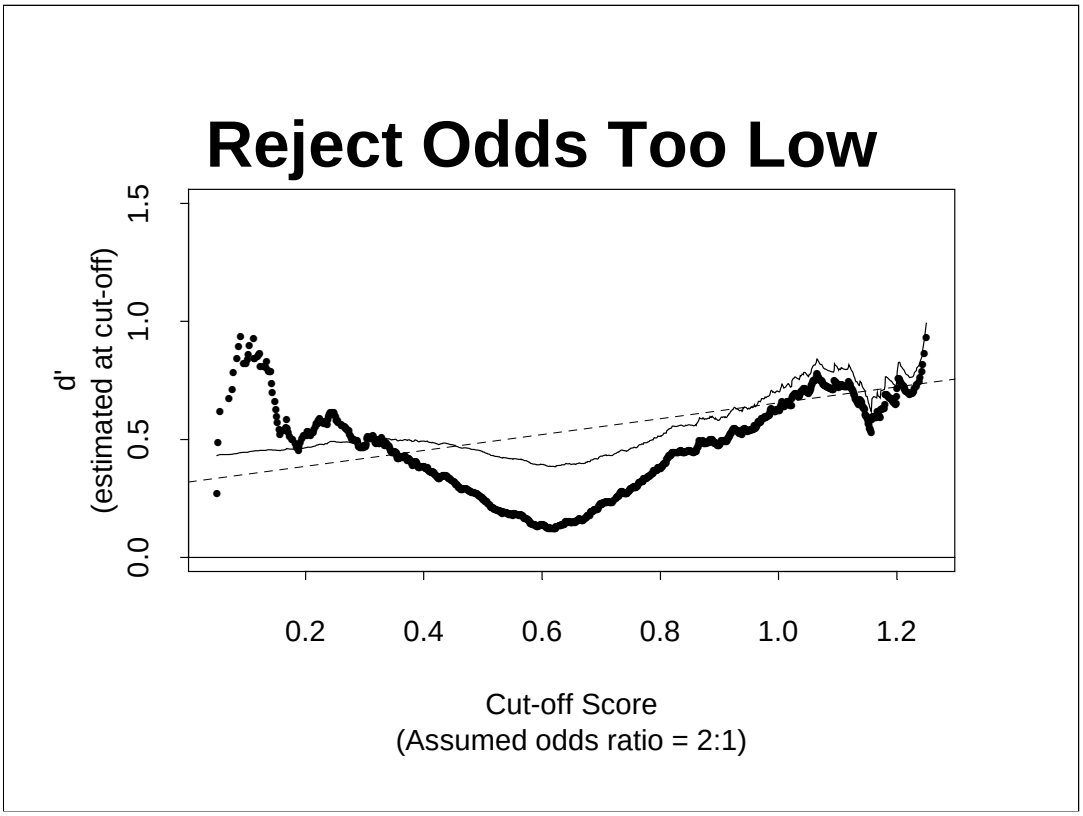
Estimating Reject Odds

- Iteratively search for the reject odds that makes the *d** curve have zero slope
- Reject odds can be added in as constants to the the cumulative distributions of accepts
- Not precise enough to rely on
  - *d** curves not guaranteed to be typical
  - Relationship of reject odds to *d** slope is flat
- Useful as a sanity check

We can't directly use the *d** observed in the accepts to determine the ratio of the accept odds to reject odds. So search for a value of reject odds that makes the *d** curve of the accepts most closely resemble the assumed form (zero slope).

If the previous acceptance decision was based on score the rejects should all fall below the bottom of the accept distributions. Therefore, their assumed odds can be added in as constants to the cumulative distributions. The search for odds of the rejects can be done before inferring reject performance at the per case level.

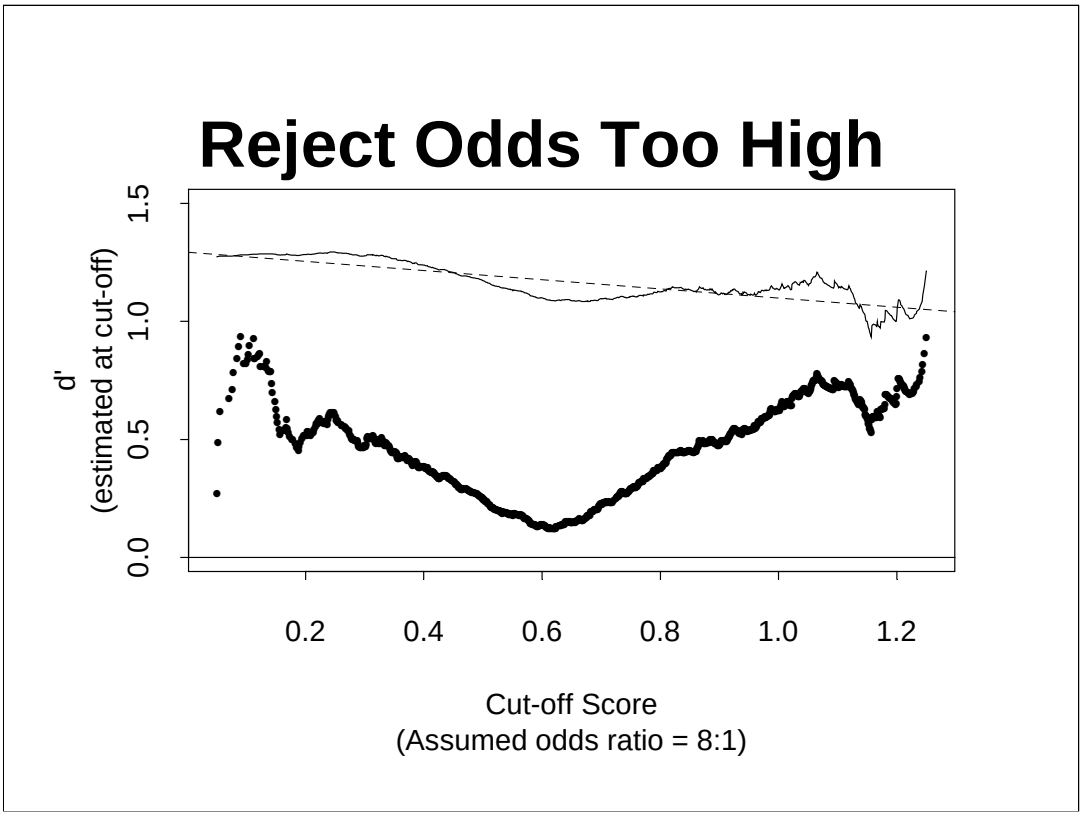
Unfortunately, this technique is not sufficiently precise to be used as the sole method for determining reject odds. The *d** curves of the accepts are not guaranteed to behave typically. The dependence of *d** slope on reject odds is not sharply peaked, so it is not possible to precisely determine the best value of the reject odds. However, the technique is useful as a check on other methods of assigning reject odds.



The dark points (mostly overlapping to form a solid line) form the sensitivity curve of the accepts. (It is not a particularly good model, although the low sensitivity is at least partly due to the restricted range of performance in the accepts because of the removal of the rejects.)

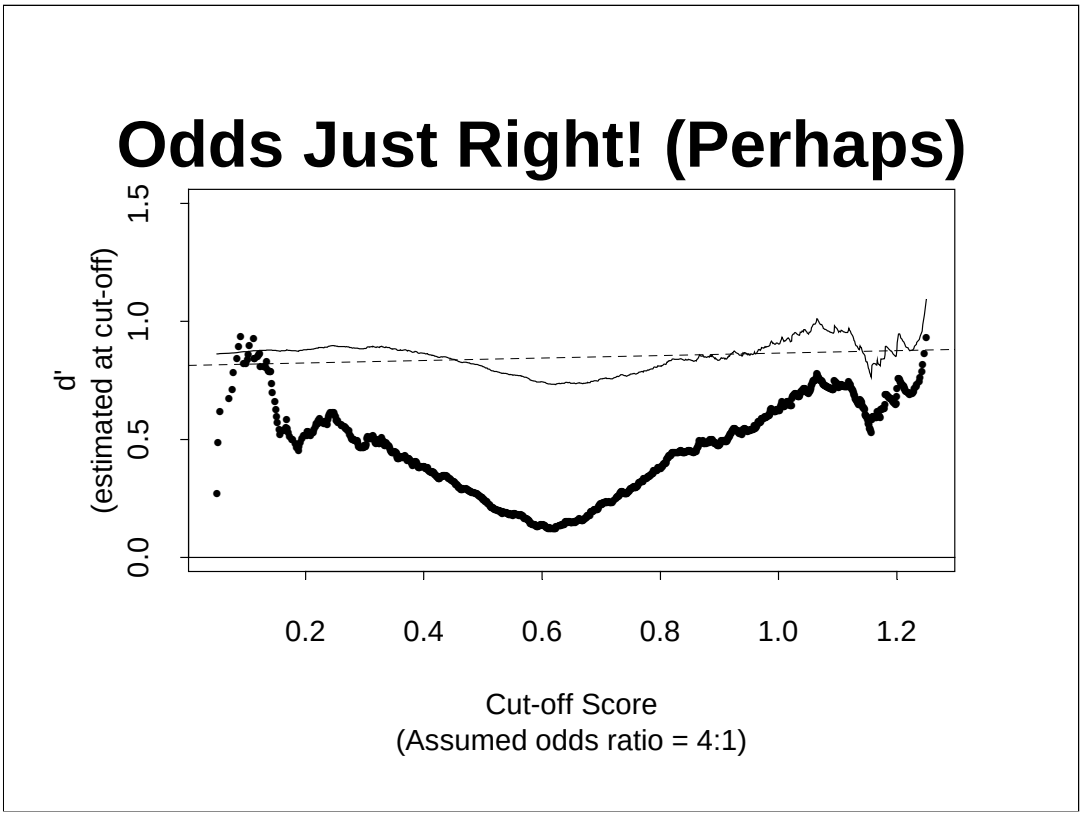
The thin solid line is the adjusted sensitivity curve after including the assumed odds of the rejects. The odds of the rejects have been set to half the odds of the accepts (a low differential relative to the heuristic).

The dashed straight line is a linear regression fit to the adjusted sensitivity curve. This line has an obvious positive slope, indicating that the ratio of the accept odds to reject odds is too low.



The odds of the rejects have been set to one eighth the odds of the accepts (a high differential relative to the heuristic).

The linear regression fit to the adjusted sensitivity curve has an obvious negative slope, indicating that the ratio of the accept odds to reject odds is too high.



The odds of the rejects have been set to one quarter the odds of the accepts (a typical differential according to the heuristic).

The linear regression fit to the adjusted sensitivity curve has almost zero slope, indicating that the ratio of the accept odds to reject odds is approximately correct (given the assumption of constant sensitivity in the population). The slopes for odds ratios of 3:1 and 5:1 are not remarkably different and would have been practically acceptable.

Sensitivity Curves in Characteristic Analysis

- Any continuous predictor can be used
  - Scores tend to respect the TSD assumptions
  - Individual characteristics less likely to
- Non-constant sensitivity is diagnostic
- Crossing or approaching the zero line suggests a nonmonotonic predictor
- Areas of decreased sensitivity prompt search for new predictors

Any continuous predictor can be used as the decision axis in constructing a sensitivity curve. Single predictive characteristics can be used.

Scores based on multiple predictors tend to respect the distributional assumptions. Individual characteristics are less likely to respect the assumptions. Therefore, sensitivity curves of individual characteristics are less likely to show constant sensitivity.

Any deviations from constant sensitivity can be used as diagnostic information for assessing and developing predictive characteristics.

A sensitivity curve that crosses or systematically approaches the zero line suggests that the predictor is not monotonically related to the outcome.

An area of decreased sensitivity should prompt the search for other predictors that are sensitive in the area of decreased sensitivity.

If I could only have one analysis I would choose to plot the log odds of outcome by predictor curve. However, the sensitivity curve is useful as a secondary analysis.

*d** and Weight of Evidence

- Weight of evidence (*W(Hypothesis:Evidence)*) is commonly used in characteristic analysis
- *d* = W(Good:>Crit.) - W(Good:<Crit.)*
- *d* = W(Good:>Crit.) + W(Bad:<Crit.)*
- *W(H:E)* is related to Bayes theorem
- log $U(H|E)$ = log $U(H) + \sum W(H:E_i)$
- Looks like a logistic regression

Characteristic analyses commonly divide continuous predictors into categories and, for each category, calculate the weight of evidence (Osteyee & Good, 1974, p. 11) as the log of the ratio of the outcome odds for the category to the outcome odds of the population. The notation for weight of evidence, *W(H:E)*, emphasises that it is the weight of evidence in favour of a hypothesis, provided by an evidence event (Good, 1960, 1968).

The calculation of *d** can be rewritten in terms of weight of evidence. It is the difference between the two evidence events in terms of the degree to which they corroborate a Good outcome. It can also be rewritten as the sum of the extent to which accepting the case corroborates a Good outcome and rejecting a case corroborates a Bad outcome (Gayler, 1988, p. 149).

Alan Turing rewrote Bayes theorem in terms of log odds and weights of evidence in order to simplify the incremental updating of the posterior odds by multiple items of evidence. Writing the prior odds of the hypothesis H as U(H) and the posterior odds of H given the evidence E as U(H|E) gives:
log $U(H|E)$ = log $U(H) + \sum W(H:E_i)$ (Good, 1982).

This has the same form as a logistic regression. The log odds of the outcome is equal to a constant plus a set of log odds increments corresponding to the predictors. The similarity is particularly apparent if the predictors are {0,1} coded dummy variables, as is often the case in credit scoring.

TSD by Regression

- Sensitivity is an increment in location along a suitably scaled decision axis
- Regression coefficients as sensitivity
- GLMs allow different transformations
  - *d'* from probit regression
  - *d** from logistic regression
- Regression formulation of TSD allows more complex models (e.g. multiple signals)

Turing's version of Bayes theorem suggests that sensitivity/discriminability can be interpreted as an increment in location of distributions on a suitably scaled decision axis. This can be cast as a regression problem (DeCarlo, 1998; Gayler, 1988, pp. 132-134; Macmillan & Creelman, 1991, pp. 281-283) so that the regression coefficients correspond to sensitivity measures.

Generalised Linear Models allow regression with a choice of link functions. These link functions correspond to different distributions and scalings of the decision axis. Probit regression yields regression coefficients scaled as *d'* and logistic regression yields regression coefficients scaled as *d** (Gayler, 1988, pp. 148-150). Other link functions are possible, for example DeCarlo examines the extreme value distribution.

Recasting TSD as a regression problem allows more complex models such as multiple and composite signals (Gayler, 1988, pp. 134-146). Other extensions are discussed by DeCarlo.

Scaling for Comparability

- Characteristic analysis gives weight of evidence on a log odds scale
- Logistic regression coefficients are sensitivities on a log odds scale
- Comparable scaling is helpful
- My favourite scale: $points = 100 \log_2 odds$
- Characteristic analysis interpreted as a scorecard with one predictor

Characteristic analysis gives the weight of evidence for each category. This is measured on a log odds scale. Logistic regression coefficients are sensitivities which are also on a log odds scale.

It is helpful to have the characteristic analyses and regression coefficients on the same scale for comparability. My favourite scale is to take 100 times the log to base 2 of the odds ratio. The base is chosen because the powers of 2 are easily remembered and $2^{1/2}$ is about the smallest odds increment of practical interest. The factor of 100 gives scores that typically fall in the range 0 to 1000 which is standard for some vendors and mandatory for some software.

This scaling allows a characteristic analysis to be interpreted as a scorecard with one predictor. The "point allocations" from the characteristic analysis may be directly compared with the corresponding point allocations from the regression model.

References

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). London: Chapman and Hall.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3,* 186-205.

Gayler, R. W. (1988). *Development of a methodology and theoretical framework for melodic discrimination*. Doctoral dissertation, University of Queensland, Brisbane, Australia. (Available from www.umi.com, University Microfilms International No. 8904966).

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society: Series B, 22,* 319-331.

Good, I. J. (1968). Corrigendum: Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society: Series B, 30,* 203.

Good, I. J. (1982). Bayes, Turing and the logic of corroboration. In D. Michie (Ed.), *Machine intelligence and related topics*. New York: Gordon & Breach.

Hand, D. J. (1997). *Construction and assessment of classification rules*. Chichester, UK: Wiley.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

Norman, D. A., & Wickelgren, W. A. (1965). Short-term recognition memory for single digits and pairs of digits. *Journal of Experimental Psychology,* 70, 479-489.

Osteyee, D. B., & Good, I. J. (1974). *Information, weight of evidence, the singularity between probability measures and signal detection*. Berlin: Springer-Verlag.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, PGIT-4, 171-212.

SAS Institute Inc. (1989) *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 2*. Cary, NC: SAS Institute Inc.

Scott, M. J. J., Niranjan, M., Melvin, D. G., & Prager, R. W. (1998) *Maximum realisable performance: A principled method for enhancing performance by using multiple classifiers* (Technical Report CUED/F-INFENG/TR. 320). Cambridge, UK: Cambridge University, Engineering Department.

Swets, J. A., & Pickett, R. M. (1982) *Evaluation of diagnostic systems: Methods from Signal Detection Theory*. New York: Academic Press.

Tanner, W. P., Jr., & Swets, J. A., (1954). A decision-making theory of visual perception. *Psychological Review*, 61, 401-409.

Wald, A. (1950). *Statistical decision functions*. New York: Wiley.

Wilkie, A. D. (1992). Measures for comparing scoring systems. In L. C. Thomas, J. N. Crook, & D. B. Edelman (Eds.) *Credit Scoring and Credit Control* (pp. 123-138). Oxford: Clarendon Press.