How to evaluate credit scorecards - and why using the Gini coefficient has cost you money

David J. Hand

Imperial College London
Quantitative Financial Risk Management Centre

August 2009

A scorecard produces a measurement of a property

Scorecards have many uses:

- compare score with threshold, to make a classification
  e.g. predict default / not default
  - predict what will happen; decide what action to take

- monitor changing values for management
  e.g. credit limit increase or decrease if score drops too much

- monitor one's own credit score
  - decide which bank to go to for a loan

How to evaluate a scorecard

depends on what it is being used for

This talk is about using scorecards for ***predictive classification***

e.g. using scorecards to predict the likely future class of a customer, e.g. default / not default

If you are using the scorecard for some other purpose, then the arguments below may not apply

Scorecards for classification: problem structure

Past	Present	Future

Design sample

Measured characteristics	➔	Outcome

Applied to

Measured characteristics	➔	Outcome

Scorecards for classification

Notation: measured characteristics *x*, outcome *c*

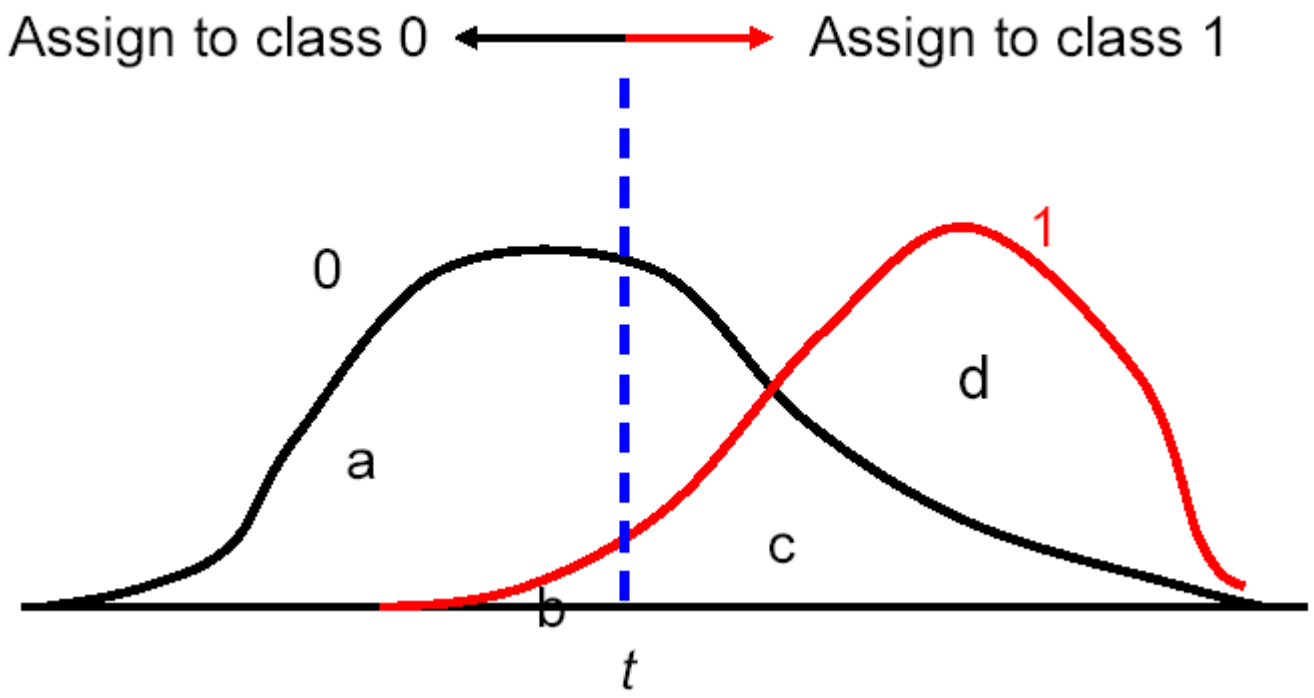
(Here assume two classes, $c = 0,1$)

Aim: use design sample $\left(x_i, c_i\right) \quad i = 1,...,n$

- construct *score* $s = f\left(x\right)$
- compare score with threshold *t*
- assign to class 0 if $s \le t$ class 1 if $s > t$

$\Rightarrow$ two issues

- how to construct the score
- how to choose the threshold



Distribution of class 1 scores is $f_1(s)$, with cdf $F_1(s)$

Distribution of class 0 scores is $f_0(s)$, with cdf $F_0(s)$

Methods for deriving score function

logistic regression

linear discriminant analysis

naive Bayes

segmented scorecards

e.g. logistic regression trees

Other methods: quadratic discriminant analysis, regularised discriminant analysis, perceptrons, neural networks, radial basis function methods, vector quantization methods, nearest neighbour and kernel nonparametric methods, tree classifiers such as CART and C4.5, support vector machines, rule-based methods, random forests, etc. etc. etc.

How to choose between them

- compare performance
- need a performance criterion

- constructing rules means estimating parameters
- do this by optimising a performance criterion

How to measure performance

Basic issues

Business vs statistical criteria

Design (training) set and test set: apparent performance

Symmetric and asymmetric problems
- in business most are asymmetric
  e.g. good/bad risk; profitable/not; fraud/legit; etc

Here let class 1 = ‘good’

Basic misclassification table

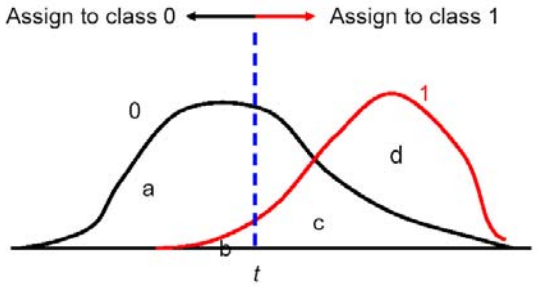
		True class	
		0	**1**
Predicted class	**0**	a	b
	1	c	d

$a+b+c+d=n$ $\quad a+c=n_0$ $\quad b+d=n_1$

Class *priors*

Proportion in class 0 = $\pi_0 = n_0/n$

Proportion in class 1 = $\pi_1 = n_1/n$



		True class	
		0	**1**
Predicted class	**0**	a	b
	1	c	d

C = Proportion correctly classified = $(a+d)/n$

E = Proportion misclassified = $(c+b)/n$

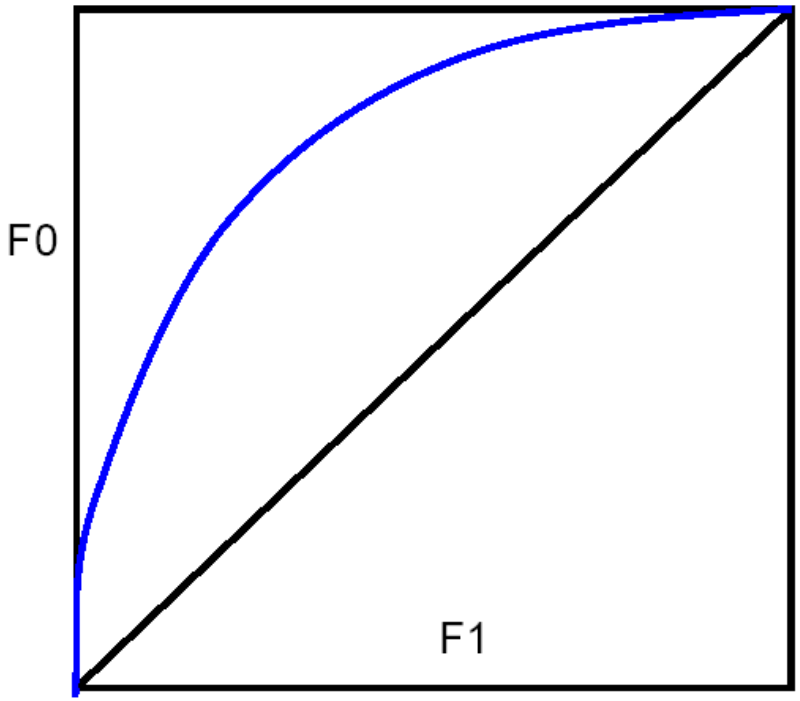
$1-F_1(t)$ = prop of class 1 correct = $d/(b+d)$

$F_0(t)$ = prop of class 0 correct = $a/(a+c)$

Br = Bad rate amongst 'accepts' = $c/(c+d)$

Gr = Good rate amongst 'rejects' = $b/(a+b)$

Can represent performance with all choices of t simultaneously with an ROC curve



The Gini coefficient is twice area between curve and diagonal
The AUC is the area under the curve: $G = 2AUC - 1$

How to choose threshold, *t* ?

Another way of looking at things

Let cost of misclassifying a class *i* case be c_i

Then overall loss due to misclassifications is

$$L = c_0 \pi_0 \left(1 - F_0\left(t\right)\right) + c_1 \pi_1 F_1\left(t\right)$$

The threshold *T* which minimises the loss is given by

$$T(c_0, c_1) \triangleq \arg\min_t \left\{ c_0 \pi_0 \left(1 - F_0(t)\right) + c_1 \pi_1 F_1(t) \right\}$$

Leading to a mapping between c_1/c_0 and *T*

$$c_1/c_0 = \pi_0 f_0(T) / \pi_1 f_1(T)$$

$\Rightarrow$ ***choosing threshold ≡ choosing*** c_1/c_0

But how can we choose the costs ?

Strategy 1:

Make default assumptions

Example 1: $c_0 = c_1 = 1$

Leading to ***error rate***

Example 2: $c_0 = \pi_1, \;\; c_1 = \pi_0$

Leading to ***Kolmogorov-Smirnov statistic***

BUT

Costs should be chosen on the basis of the problem, on business grounds

Not picked for mathematical convenience !

Should you be choosing customers just to make the maths easy?

Strategy 2:

Average over all possible costs

$$L = \int_0^{\infty}\int_0^{\infty} \left\{ c_0 \pi_0 \left(1 - F_0\left(T(\mathbf{c})\right)\right) + c_1 \pi_1 F_1\left(T(\mathbf{c})\right) \right\} u(c_0, c_1)\, dc_0 dc_1$$

$$L = \int_0^{1} \left\{ c \pi_0 \left(1 - F_0\left(T(c)\right)\right) + (1-c) \pi_1 F_1\left(T(c)\right) \right\} w(c)\, dc$$

with

$$c = \left(1 + c_1/c_0\right)^{-1} \quad \text{and} \quad c = \pi_1 f_1(T) \Big/ \left(\pi_0 f_0(T) + \pi_1 f_1(T)\right)$$

A particular choice of $w(c)$ in

$$L = \int_0^1 \left\{ c\pi_0 \left(1 - F_0\left(T(c)\right)\right) + (1-c)\pi_1 F_1\left(T(c)\right) \right\} w(c)\, dc$$

leads to

$$L = 1 - 2\pi_0\pi_1 \int_{-\infty}^{\infty} F_0(s)\, f_1(s)\, ds$$

$$= 1 - 2\pi_0\pi_1 AUC \qquad = (1 - \pi_0\pi_1) - \pi_0\pi_1 G$$

This choice is

$$w^*(c) = \pi_0 f_0\left(T(c)\right)\frac{dT(c)}{dc} + \pi_1 f_1\left(T(c)\right)\frac{dT(c)}{dc}$$

That is, AUC and Gini are equivalent to averaging the loss over a weight function $w(c)$ ***which depends on the observed distributions***

BUT

Knowledge of c, or, equivalently, knowledge of T or the costs c_0 ***and*** c_1 ***, must come from information in the problem other than the score distributions !***

Using the *AUC* or *Gini* is equivalent to saying your belief in the relative severity which will be assigned to the different types of misclassifications depends on choice of scorecard

This is absurd

Like measuring my height in millimetres, and yours in feet, and saying I am taller because my number is larger than yours

But (you say) the AUC and Gini have several nice interpretations

e.g. *AUC* is the average value of $F_0\left(s\right)$ if $v = F_1\left(s\right)$ is chosen from a uniform $x \sim \left[0,1\right]$ distribution

$$AUC = \int_0^1 F_0\left(F_1^{-1}\left(v\right)\right)dv$$

F1 and F0 are *operating characteristics* of the scorecard

I can *choose* F1 (or F0) according to whim

In particular, I might choose a different 1-F1 for different rules

	t		**t'**	
	1-F1(t)	**F0(t)**	**1-F1(t')**	**F0(t')**
Rule 1	95	89	80	90
Rule 2	95	10	80	90

- In contrast, the distribution $w(c)$ is a matter of *belief*, not choice
- It measures the beliefs about the relative severity of the different kinds of misclassification
- $w(c)$ cannot vary from scorecard to scorecard
- **I cannot say**: *if I use logistic regression then misclassifying a good as a bad is ten times as serious as the reverse,*
  but if I use linear discriminant analysis then it is a hundred times as serious
- c ***is a property of the problem, not the scorecard***

Relationship between costs and sensitivity 1-F1 depends on empirical score distributions f_0 *and* f_1

$\Rightarrow$ fundamental ***complementarity***

$\varphi(v)$ ~ choice of *x* distribution for the averaging $1-F_1$

$w(c)$ ~ choice of *c* distribution for the averaging costs

One cannot simultaneously choose φ and *w* independently of the empirical distributions

If one distribution is independent of the empirical distributions, the other distribution necessarily depends on the classification rule used

Which should we use?

v is a matter of ***choice***, *c* is a ***property of the problem***

AUC and *Gini* use $\varphi(v) \sim U[0,1]$

$\Rightarrow$ ***AUC and Gini average over cost distributions which vary from scorecard to scorecard***

[Same applies to Partial AUC]

*It is OK for **different** people to choose **different** w*

- because they are interested in different aspects of performance

*It is wrong for a **single** person to choose **different** w for **different** scorecards*

- doing so implies using different measures for different scorecards

Conclusion

- The Gini coefficient does not compare like with like
- The Gini coefficient can lead to mistaken performance comparisons
- The Gini coefficient can result in incorrect decisions

What to do instead?

Need to choose a weight function which is the same for all scorecards

Choice (1): choose *w* to reflect your personal beliefs in the likely values of *c* (for each problem)

Choice (2): choose a universal standard *w*

- I suggest $beta(2,2)$

Recommendation: *use both*

Alternative measure

Choose weight function invariant to choice of scorecard:

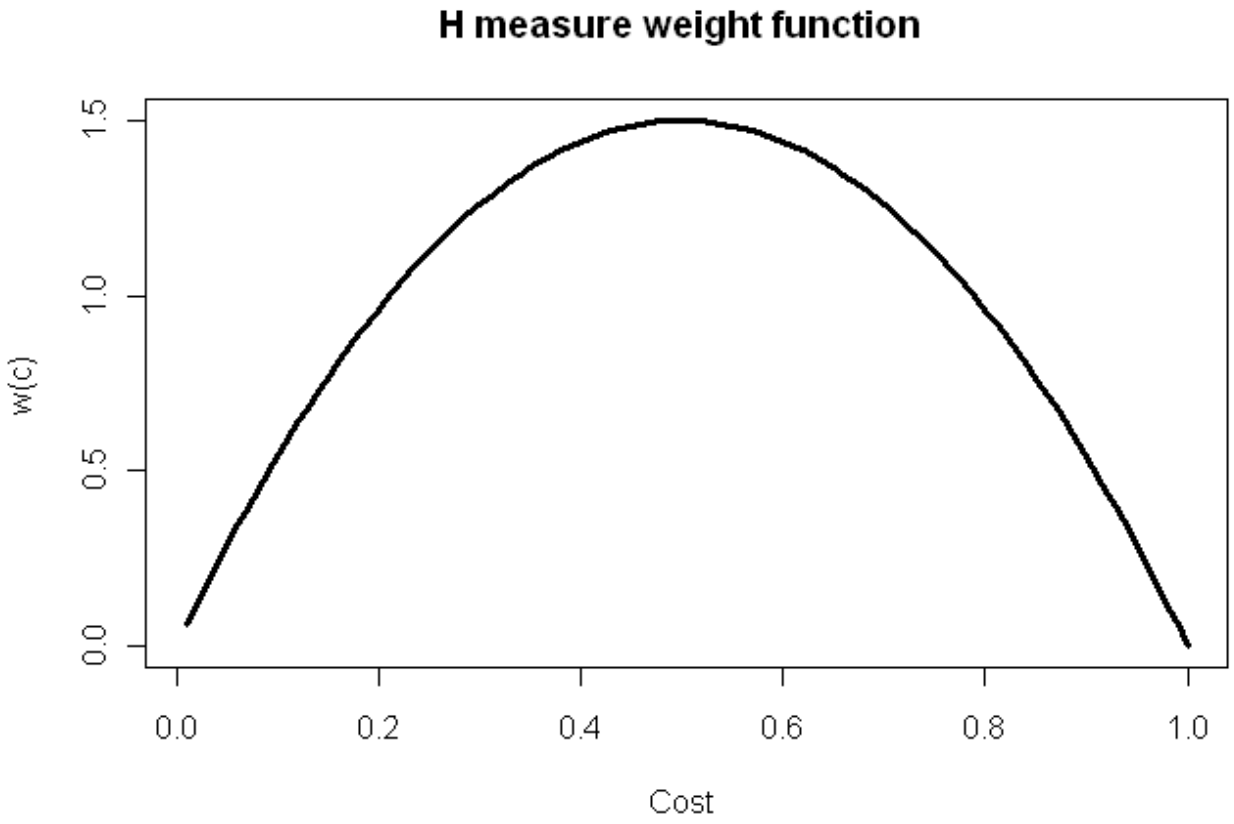
$$w(c) = w_{\alpha,\beta}(c) = \frac{c^{\alpha-1}(1-c)^{\beta-1}}{B(1;\alpha,\beta)}$$

Leading to the *H* measure

Standardise

- to lie between 0 and 1
- with large value being good

Standard *H*: with $(\alpha, \beta) = (2, 2)$



Example *(but not needed !)*:

Whether or not someone will settle outstanding debts immediately: 9 predictors, 6378 cases

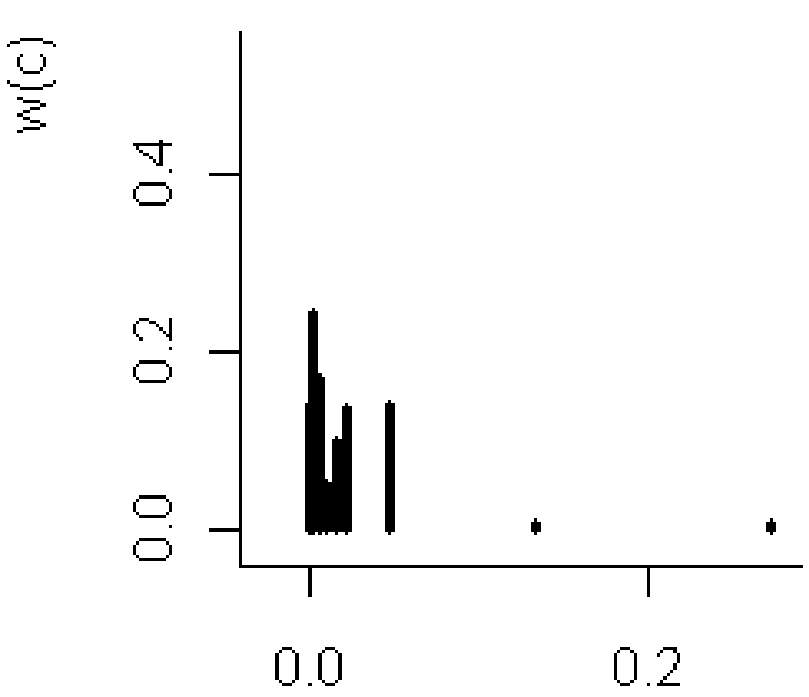
Classifier (i): ***linear discriminant analysis***

Classifier (ii): ***logistic regression***

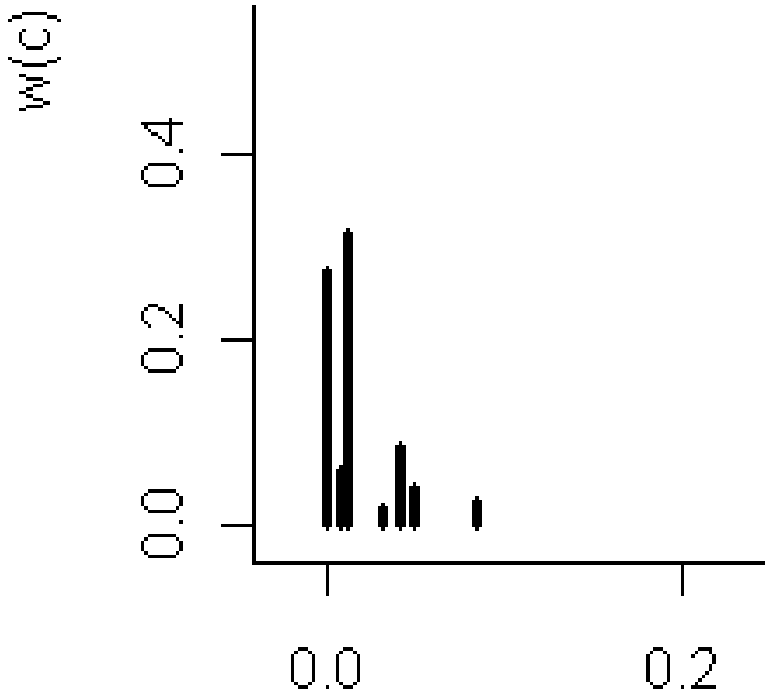
	Classifier (i)	**Classifier (ii)**
Gini	**0.580**	**0.500**
H	**0.016**	**0.027**

The Gini coefficient uses different $w^*(c)$ ***distributions for the classifiers:***

Linear discriminant



Logistic regression



Some references:

Hand D.J. (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. To appear in *Machine Learning*.

R code for the *H* measure is available on http://stats.ma.ic.ac.uk/d/djhand/public_html/

Krzanowski W.J. and Hand D.J. (2009) *ROC curves for continuous data*. CRC Press.