Assessing and evaluating scorecards

Tools, methods, and consequences

David J. Hand
Imperial College
and
Winton Capital Management

30 August 2013

Scorecards are statistical models for evaluating *risk*

They are used for multiple purposes, including:

- to decide who should receive a financial product
- to protect applicants from overstretch
- to decide interest rates
- to detect fraud
- to guide customer management
- to make profit for the company
- to increase shareholder value
- ....

Simplest (almost generic) scenario:

predict who will default

just two classes, labelled ‘good’ and ‘bad’

Two stages

1) compute credit *score*
2) choose decision *threshold*

*Modern credit scoring is '**data-driven**': based on observed relationships in the data*

a statistical revolution

*Contrast with '**substantive**': based on theory relating characteristics to default probability*

The fact that credit scoring is *data-driven* has sometimes worried the regulators

- who might prefer 'substantive' models based on *narrative*
  e.g. hypothesised causation
- quantitative, data-driven, modellers base on *data*
  e.g. association, with no obvious causal link
- and there are complications like legal constraints

Why data driven?

Because our aim is to ***build the best models we can***

- use *any* characteristics which improve prediction
- transform/combine in arbitrarily complex ways

Many approaches have been used

e.g. *Linear discriminant analysis, quadratic discriminant analysis, naive Bayes, regularised discriminant analysis, logistic regression, SIMCA, DASCO, logistic regression, perceptrons, neural networks, support vector machines, tree classifiers, random forests, nearest neighbour, Parzen kernel methods, ...*

But by far the most common is a *logistic regression tree*
e.g. 10 segments, each with 20 predictors

Scale typically transformed, e.g. to discrete scores on range 300-850)

why ***logistic*** ?
why ***regression*** ?
why a ***tree***?

Logistic because it gives a model of default probability
Tree because it is (equivalent to) including interactions
Regression because weighted sums easy to interpret

But ...

This can be misleading:

In a weighted sum

$$y = \text{logit}\left[P\left(good \right) \right] = w_1 x_1 + w_2 x_2 + w_3 x_3$$

(with $V\left(x_1 \right) = V\left(x_2 \right) = V\left(x_3 \right) = V\left(y \right)$ for simplicity)

the impact of x_1 on *y* is ***not*** given by the weight w_1

but by

$$\rho\left(y, x_i \right) = \rho\left(w_1 x_1 + w_2 x_2 + w_3 x_3, x_1 \right) = w_1 + w_2 \rho\left(x_1, x_2 \right) + w_3 \rho\left(x_1, x_3 \right)$$

So we can say

'... a low score on characteristic XXX (e.g. prior credit record) is associated with a poor score...'

but in determining the strength of that association we must use the ***marginal*** relationship, not the ***conditional*** one

Recall (e.g.) US Equal Credit Opportunity Act:

'*applicant has the right to request the reason(s) for denial within sixty days of receipt of the creditor's notification, along with the name, address, and telephone number of the person who can provide the specific reason(s) for the adverse action*'

But perhaps most important of all

Logistic works ! - and often better than alternatives

Data set	Best method e.r.	Lindisc e.r.	Default rule	Prop linear
Segmentation	0.0140	0.083	0.760	0.907
Pima	0.1979	0.221	0.350	0.848
House-votes16	0.0270	0.046	0.386	0.948
Vehicle	0.1450	0.216	0.750	0.883
Satimage	0.0850	0.160	0.758	0.889
Heart Cleveland	0.1410	0.141	0.560	1.000
Splice	0.0330	0.057	0.475	0.945
Waveform21	0.0035	0.004	0.667	0.999
Led7	0.2650	0.265	0.900	1.000
Breast Wisconsin	0.0260	0.038	0.345	0.963

Prop linear = (Default - LDA) / (Default - Best)

General comments about *assessing* scorecards

- used for evaluating (to ask: is it good enough?)
- for comparing (to ask: this one or that one?)
- for developing (e.g. does adding another characteristic improve the scorecard; does a transformation help?)

Many *domain specific* issues which need evaluation:
- simplicity of interpretation
- simplicity of construction
- ease of updating
- propensity for chosen characteristics to have missing values
- robustness (e.g. to out of date training data)
- effectiveness of coping with a nonstationary world (sensitivity to change)
- ***and accuracy of the assignments***

***NOTE*:**

clearly no single measure can capture all these aspects

This suggests a *profile* of measures should be used

I'm going to talk about

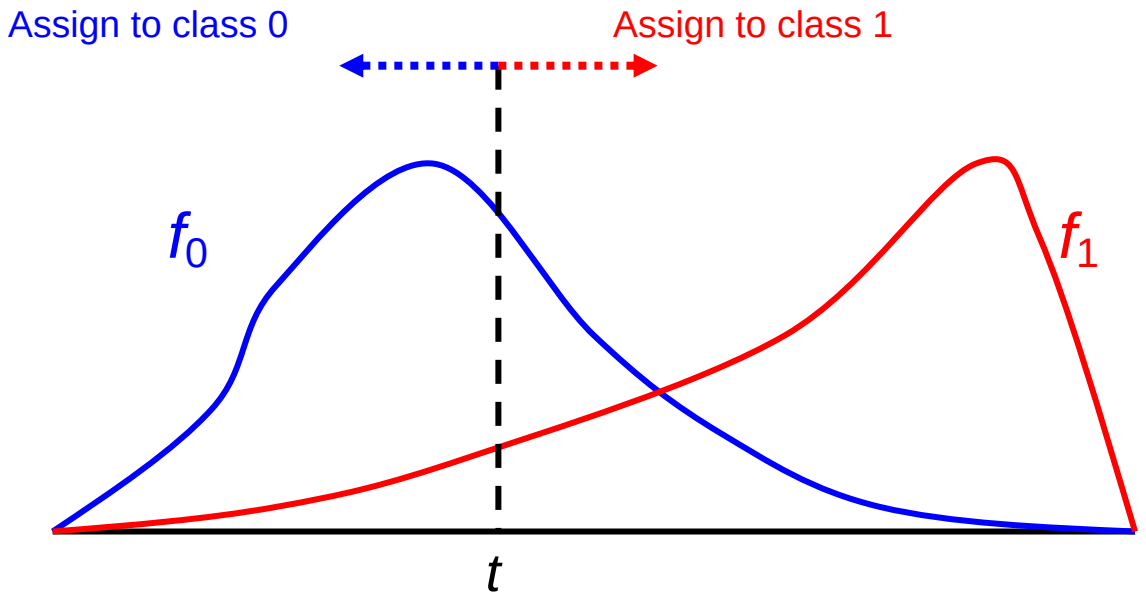
- application scoring
- *creditworthiness* scores

Concern is whether or not an applicant will default

My concern here is solely with 'accuracy' of class assignment

$f_0(s)$ = distribution of scores for class 0, cdf $F_0(s)$

$f_1(s)$ = distribution of scores for class 1, cdf $F_1(s)$

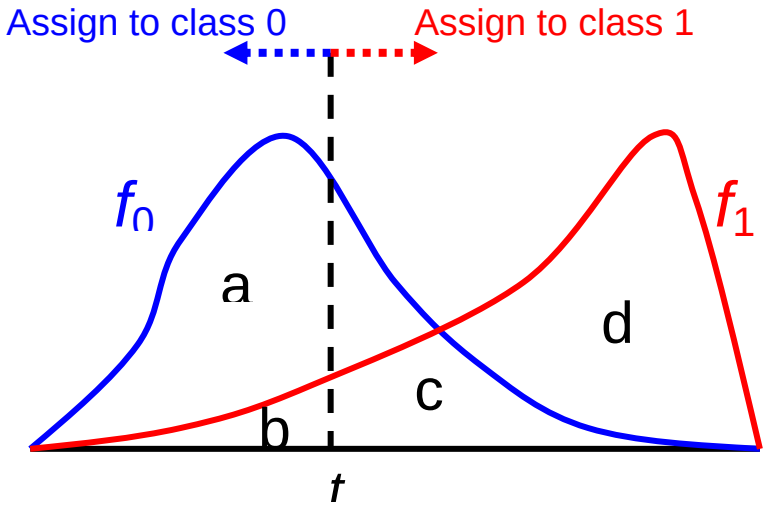


For simplicity in what follows

1) I shall assume the distributions are accurately estimated
2) I shall assume $F_0(s) > F_1(s)$ for all s

A given *t* yields a misclassification table

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



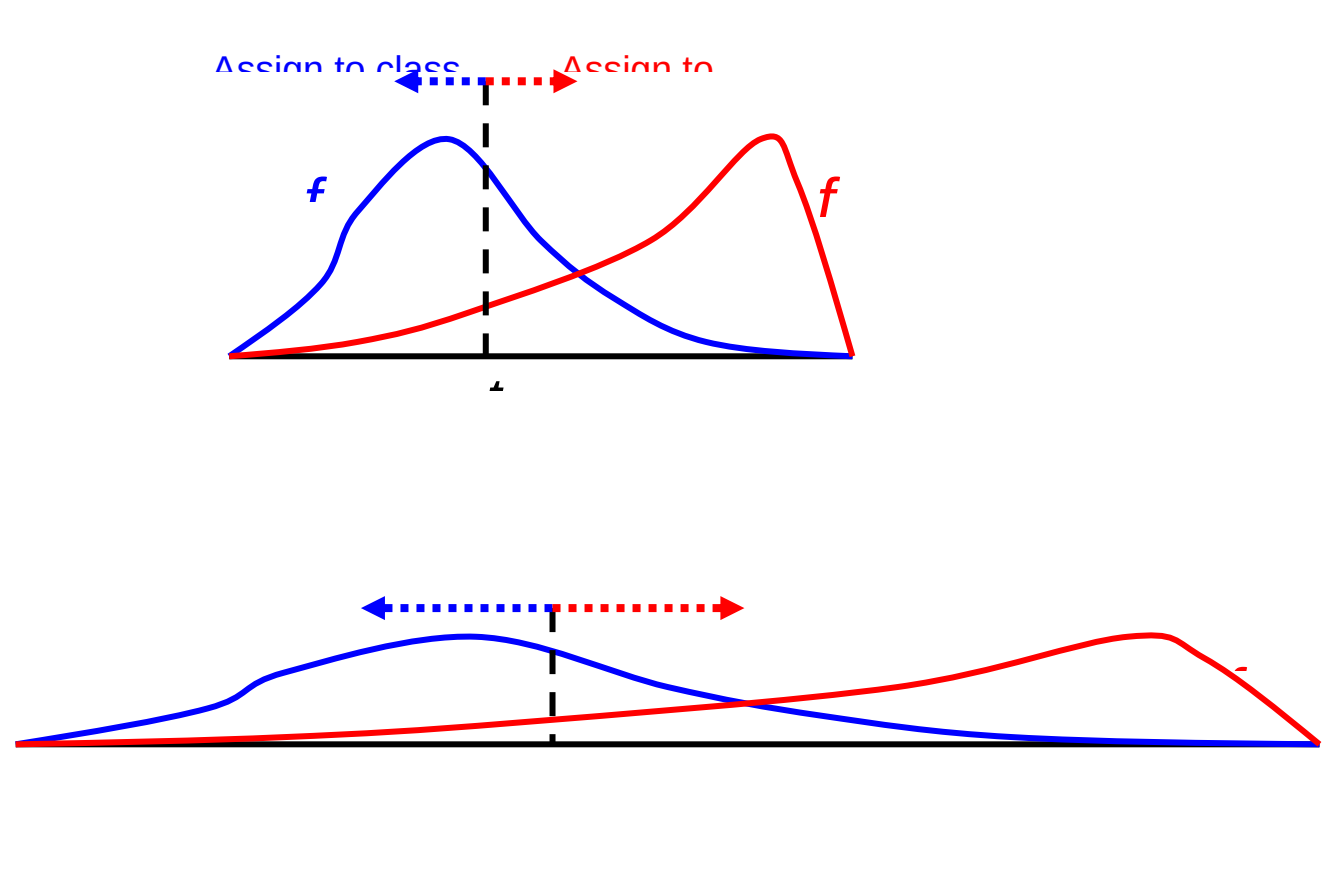
which yields various measures of performance

Note:

1) need simple numerical summary so can automatically search and compare models

2) given *t*, distance (*s* - *t*) is irrelevant, only sign matters
(could build separate model for severity of error)

This means that the measures must be invariant to monotonic transformations of the score
which is why the estimated P(1|x) can be transformed to arbitrary scales (e.g. 300-850) in scorecards


f
f
f
f
0
1
t

Calibration

Here I define a scorecard as *calibrated* if

$$\pi_1 f_1(s) / f(s) = s$$

That is:

- of applicants with score *s*, a proportion *s* belong to class 1;
- the probability is *s* that an applicant with score *s* will belong to class 1;

Note:

calibration is not classification accuracy

e.g. assign everyone score $s = \pi_1$,

so that the scorecard is perfectly calibrated

but useless for decision making

To calibrate a scorecard, estimate the probability of class 1 membership at each score and apply a monotonic transformation of estimated score so that $\pi_1 f_1(s)/f(s) = s$

Since $\pi_0 f_0(s) + \pi_1 f_1(s) = f(s)$

We have that, for calibrated scorecards,

$$f_0(s) = f(s) \times (1-s)/\pi_0 \qquad f_1(s) = f(s) \times s/\pi_1$$

[and hence, for calibrated scorecards: $f_0(s) = f_1(s) \times (1-s)\pi_1 / s\pi_0$]

And a little calculus shows $\mu = \pi_1$

and $$\sigma^2 = \pi_1(\mu_1 - \mu)$$

Henceforth suppose we have calibrated the scorecard

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

If class 1 are ‘goods’, various measures

c / (c+d)	bad rate amongst accepts [*precision* = 1 - c / (c+d) = d / (c+d)]
(c+d) / (a+b+c+d)	proportion accepted
(a+d) / (a+b+c+d)	proportion correctly classified, p_c [$p_E = 1 - p_c$, error rate]
d / (b+d)	proportion of goods correctly classified [*sensitivity*, *recall*]

and other ratios in various contexts

Kappa statistic: chance adjusted proportion correct

$$K = \frac{p_c - p_{Ch}}{1 - p_{Ch}} = \frac{2(ad - bc)}{(a+b)(a+c)+(c+d)(b+d)}$$

F measure

$$\frac{2d}{(b+c)+2d} = \frac{2}{(g.r.\ among\ acc.)^{-1} + (prop.\ g.\ corr.\ class'd)^{-1}}$$

Matthews coefficient (= Pearson correlation)

$$\frac{ad - bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$$

Ideally the threshold should be chosen on the basis of knowledge

- what is the desired (estimated) proportion to be accepted
- what is the desired (estimated) bad rate amongst accepts
- ....

Or the choice can be based on minimising some performance measure

But often (usually) the threshold, *t*, is unknown

Two strategies

STRATEGY 1: Choose *t* to optimise some criterion

e.g. 1) Proportion correctly classified

$$\max_t \left(\pi_0 F_0(t) + \pi_1 \left(1 - F_1(t)\right)\right) \qquad = \pi_1 + \max_t \left(\pi_0 F_0(t) - \pi_1 F_1(t)\right)$$

For fixed class proportions the initial π_1 is irrelevant

so this measure is equivalent to $\max_t \left(\pi_0 F_0(t) - \pi_1 F_1(t)\right)$

e.g. 2) KS (equivalent to Youden statistic)

$$\max_t \left(F_1(t) - F_0(t)\right)$$

So the KS measure is simply the proportion correctly classified if the two classes were of equal size

This could be good or bad
Depends on problem

But the choice of threshold in the KS measure depends on the shapes of $f_0(s)$ and $f_1(s)$

Since the scorecards are calibrated, this means that KS is equivalent to assigning to class 1 if the estimated class 1 prob is greater

than t_A with scorecard A
than t_B with scorecard B
where $t_A \neq t_B$

Or: there are applicants for which the scorecards give the same estimated probability of being good, but for whom one accepts and the other rejects

STRATEGY 2: Average over a distribution of possible *t*

Case 1: assign to class 1 those applicants with estimated class 1 probabilities greater than *t*

This is sensible: any alternative would mean that classifiers which agree that an object had estimated probability *p* of belonging to class 1 could assign it to different classes

We don't know *t*, so we take a distribution

Gini = 2×AUC - 1

AUC tells us what proportion of class 0 points we get right, on average when we choose *t* from the distribution $f_1(t)$

$$\text{AUC} = \int F_0(t)\, f_1(t)\, dt \qquad = 1 - \int F_1(t)\, f_0(t)\, dt = \int \left(1 - F_1(t)\right) f_0(t)\, dt$$

equivalent to the proportion of class 1 points we get right on average when we choose *t* from the distribution $f_0(t)$

But the scorecards are calibrated

Why would we consider different distributions of the threshold for different classifiers?

Taking different distributions is equivalent to saying that, if logistic regression were to be used, then we would be very likely to choose probability 0.9 as our classification threshold, whereas if a random forests classifier were to be used we would be very *unlikely* to choose probability threshold 0.9

The threshold distribution should be a property of the *problem*, not the classifier, and so should be the same for all classifiers applied to the same problem.

Another way of looking at this

$$AUC = \int F_0(t)\, f_1(t)\, dt$$

$$= \frac{1}{\pi_0} \int \left[\pi_0 F_0(t) + \pi_1 \left(1 - F_1(t)\right) \right] f_1(t)\, dt - \frac{\pi_1}{2\pi_0}$$

$$= \frac{1}{\pi_0} \int M(t)\, f_1(t)\, dt - \frac{\pi_1}{2\pi_0}$$

where $M(t) \triangleq \pi_0 F_0(t) + \pi_1 \left(1 - F_1(t)\right)$, the overall proportion correctly classified when threshold *t* is chosen

So AUC and Gini are equivalent to an *average proportion of applicants correctly classified*, as the threshold varies

But where the distribution taken for the threshold varies between scorecards

$$Gini = \frac{2}{\pi_0}\int M(t)\, f_1(t)\, dt - \frac{1}{\pi_0}$$
$$= a + b \times \left[mean \;\; proportion\; correctly\; classified \right]$$

(where *a* and *b* are functions of the class sizes only)

$\Rightarrow$ *Gini is a silly measure*

Because the measuring instrument depends on the thing being measured

Case 2: assign a given proportion of applicants to class 1

Sometimes additional information should be taken into account

For example:
- we might want to accept a proportion *Q* of applicants
  (perhaps because this is determined by our available funds)

This is equivalent to rejecting a proportion *P = 1 - Q* of applicants

Given the (estimated) overall score distribution, for any *P* we know the appropriate threshold *t*, by inverting

$$P = F(t) = \pi_0 F_0(t) + \pi_1 F_1(t)$$

to give $t = G(P)$

Now

$$Gini = 2\int \left(1 - F_1\left(t\right)\right) f_0\left(t\right) dt - 1$$

$$= 2\pi_0^{-1} \int \left(1 - F_1\left(t\right)\right)\left[\pi_0 f_0\left(t\right) + \pi_1 f_1\left(t\right)\right] dt - 1 + \pi_1\pi_0^{-1}$$

$$= a + b \times \int \left(1 - F_1\left(G\left(P\right)\right)\right) dP$$

So using the *Gini* is equivalent to saying you think it is equally likely that you will want to accept any proportion of the population

You think it equally likely that you will want to make loans to 99% of the applicants or just 1%

This is unrealistic: So the Gini (and AUC) are inappropriate measures in this case also

Distinction between Case 1 and Case 2

In case 1 only relevant information for each applicant is their score and the threshold

In case 2, also concerned about the scores of the *other* applicants

But both cases lead to unrealistic evaluations

Gini (and AUC) is not measuring anything of interest

There are many relationships between AUC (equiv Gini), KS, proportion correct (equiv error rate) and other measures which are easy to derive by geometric arguments

I'll just mention a couple here

Simplest is Gini $\geq$ KS

Let $C(t)$ be the proportion correctly classified at threshold *t*

Then $$Gini = \left(2\int C(t)\,dF_1(t) - 1\right) \Big/ \pi_0$$

Compare with $$Gini = \left(2\int F_0(t)\,dF_1(t) - 1\right)$$

Fundamental property of AUC, Gini, KS - not dep on class sizes

Don’t forget the reject distortion

What are you actually trying to measure?

- the performance of a scorecard built using cases you previously thought merited acceptance
- the performance of a scorecard aimed at the population of applicants

Conclusions

- Use several measures
- Use a measure which matches the problem
- Dangers of poor decisions if use poor measures
- Adopt the right measures when developing scorecards

- Gini and KS are generally inappropriate

But if you must use a single performance measure

Use ***average bad rate amongst accepts***

averaged over a predetermined distribution, g*, of accept proportions (not a function of* f_0 *or* f_1*)*

$$\int [bad\ rate\ among\ accepts\ at\ thresold\ P]\, g(P)\, dP$$

$$= \int \left[\frac{\pi_0 \left(1 - F_0(P)\right)}{\pi_0 \left(1 - F_0(P)\right) + \pi_1 \left(1 - F_1(P)\right)} \right] g(P)\, dP$$

thanks!

AUC (Gini) vs the H-measure

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