Learning in the real world

David J. Hand

Imperial College London
and
Winton Capital Management

In science:

first develop the big idea
then handle the details
because the world is a complex place

so that theory ***does not quite*** match observations

For example:

Newton's laws of motion

- many interacting bodies - too much detail
- relativity - so Newton's laws are an approximation

The ‘details’ in learning theory include things like

Data quality

Missing data

Measuring the right thing

Uncertainty and ambiguity about objectives

Non-stationarity

Reactive non-stationarity

How to measure performance

etc

The ‘details’ arise from the context

All problems have different contexts

All problems are different

For example:

Often some data are ***missing***

Missing *fields* in records: obvious potential problems

Missing *records*: *you don't see what you don't see*

Not always obvious that something is missing

A vehicle to illustrate the ideas:

The comparison of classification rules

Two-class classification abstract structure:

*Given a set of objects, for each of which we know their true class and also a vector of descriptive variables, derive a rule which will allow one to classify new objects from their descriptive vectors **as accurately as possible***

Function mapping the descriptive vector to a 'score' s

Threshold t such that $s > t$ ➔ assign to class 1

$s \leq t$ ➔ assign to class 0

Basic structure:

Build classifiers using training data

Apply classifiers using test data

See which is best

The details here:

what does ‘***best***’ mean

what does ‘***as accurately as possible***‘ mean?

Problem-based criteria

vs

Classification accuracy criteria

Problem-based criteria

- speed of construction
- speed of classification
- ability to handle very large data sets
- effectiveness on small-n-large-p problems
- ability to cope with incomplete data
- interpretability
- ease with which important features can be identified
- unbalanced data sets
- accuracy of probability estimates
- etc

Classification accuracy criteria

- Sensitivity (recall), specificity
- Positive predictive value (precision)
- Negative predictive value
- Error rate
- Kappa statistic
- F-measure
- Youden statistic, KS, maximum vertical distance, ...
- AUC **(!)**
- H-measure

etc

A case study: Comparing credit scorecards

Descriptive vectors:
- applicant characteristics, past credit behaviour
- can be very high dimensional (10s of thousands)

Can be very large data sets (millions, billions)

Score may be used for classification and decision making

Example 1: Building a new (better) scorecard

Selection bias: a fundamental problem

Training set:

Existing customers, with known characteristics,
and known ‘good/bad’ outcomes

But ‘existing customers’ are those we previously thought were likely to be good

They are not a random sample from the population of potential applicants

Extreme illustration:

Binary feature X: highly predictive

X = 1 ➔ will certainly default

X = 0 ➔ will certainly not default

All other predictors, Y, are poor

So we previously rejected all those applicants with X=1

Our training set contains *none* with X=1

So when we build a new classifier, variable X is not identified as a good predictor

and we are left only with the poor predictors Y to use in our new scorecard

To tackle use ‘***reject inference***’

= attempt to infer true class of the previously rejected applicants

- reweight good/bad proportions at each predictor vector
- extrapolate the estimated probability of being bad
- accept a sample of those who should be rejected
- follow up the rejects with other banks
- etc

= need extra information from somewhere

Example 2: Choosing a new scorecard

Changing economic conditions
Changing competitive environment
Changing financial products

Mean that *scorecard performance degrades over time*
So scorecards need to be updated and replaced

Comparison is central to this process

Vignette: *Startup scoring company TopScore claims its new SVM scorecard is superior to the current neural network scorecard and a test is set up*

‘Old’ scorecard = existing ANN one
‘New’ scorecard = TopScore SVM

Test format: apply both scorecards to the same sample of customers and see which gives best results

But this sample of customers will have been accepted using the old scorecard:

➔ **a fundamental data asymmetry**

With implications ....

$f\left(s_o, s_n\right)$ is the joint population density function of old and new scores for the ***bad*** class

with marginal density functions $f_o\left(s_o\right)$ and $f_n\left(s_n\right)$

$g\left(s_o, s_n\right)$ is the joint population density function of old and new scores for the ***good*** class

with marginal density functions $g_o\left(s_o\right)$ and $g_n\left(s_n\right)$

Corresponding cdfs F_o, F_n, G_o, and G_b.

Assume *goods* tend to score higher than *bads*:

➔ $F_o\left(s\right) > G_o\left(s\right)$ and $F_n\left(s\right) > G_n\left(s\right)$

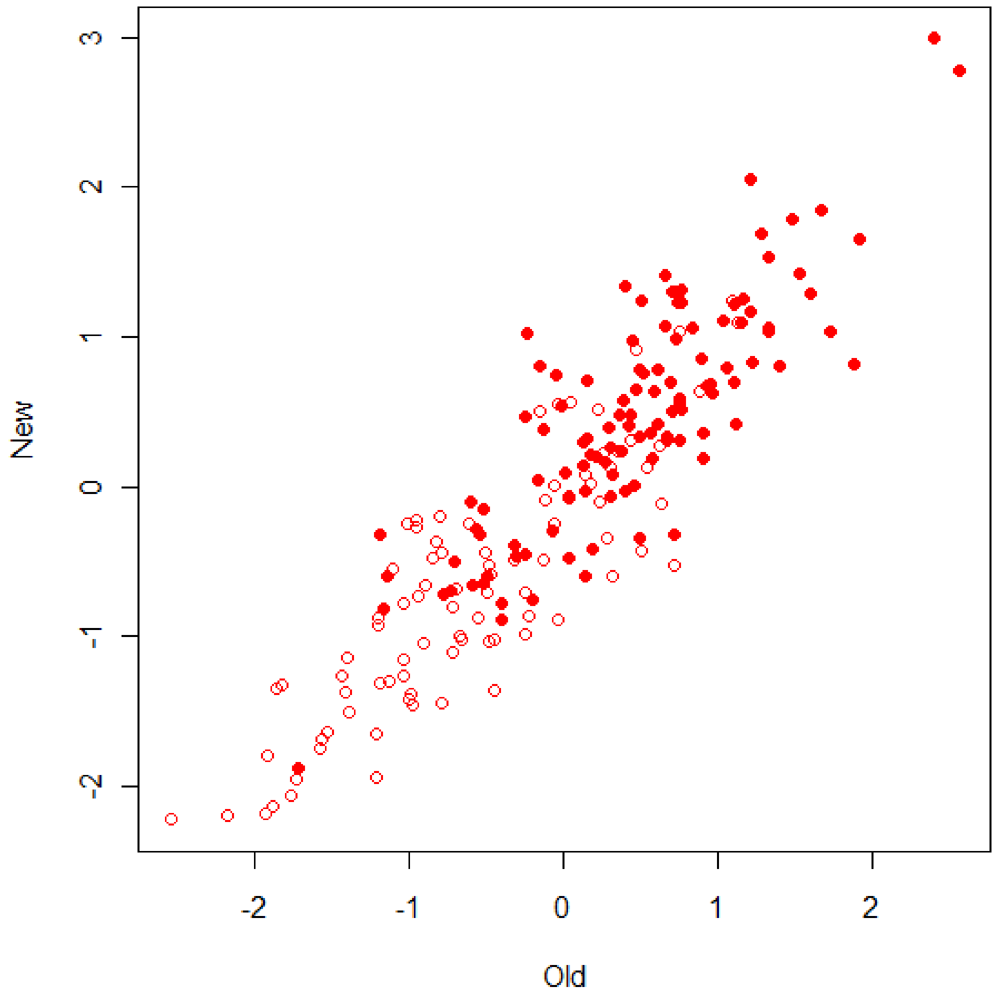
The crunch:

Training data have scores $\{(s_o, s_n) : s_o > t\}$

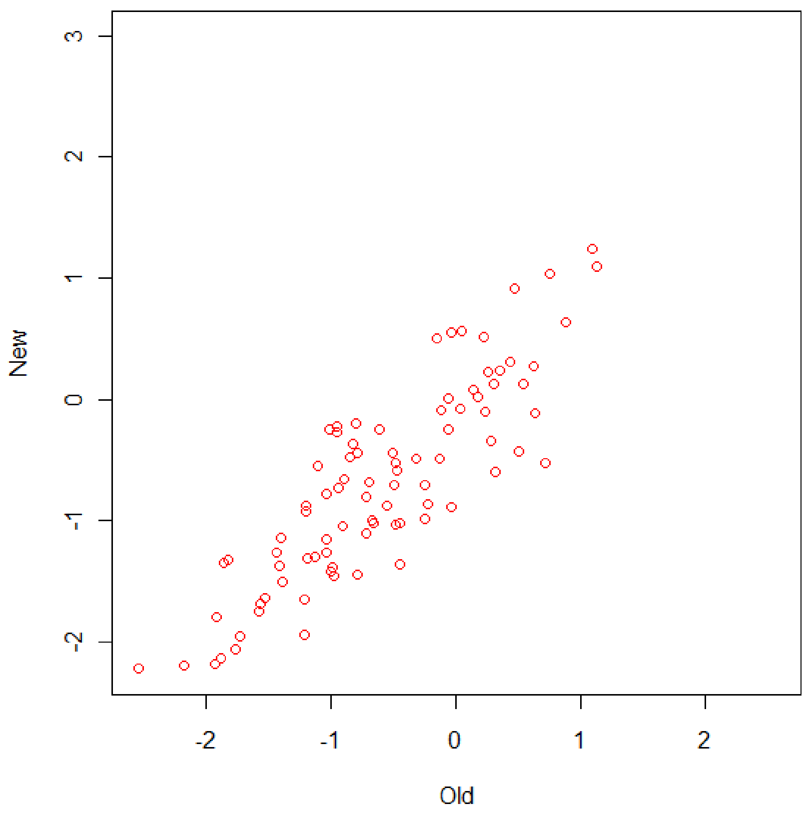
Example:

Population (s_o, s_n) bivariate normal

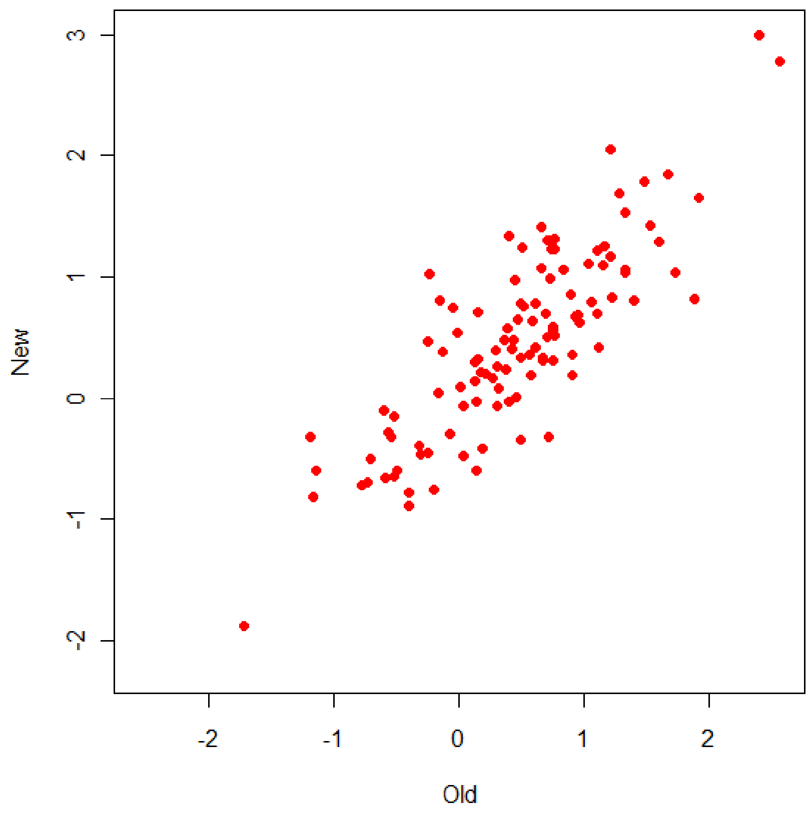
Prob of being good increases with $s_o + s_n$

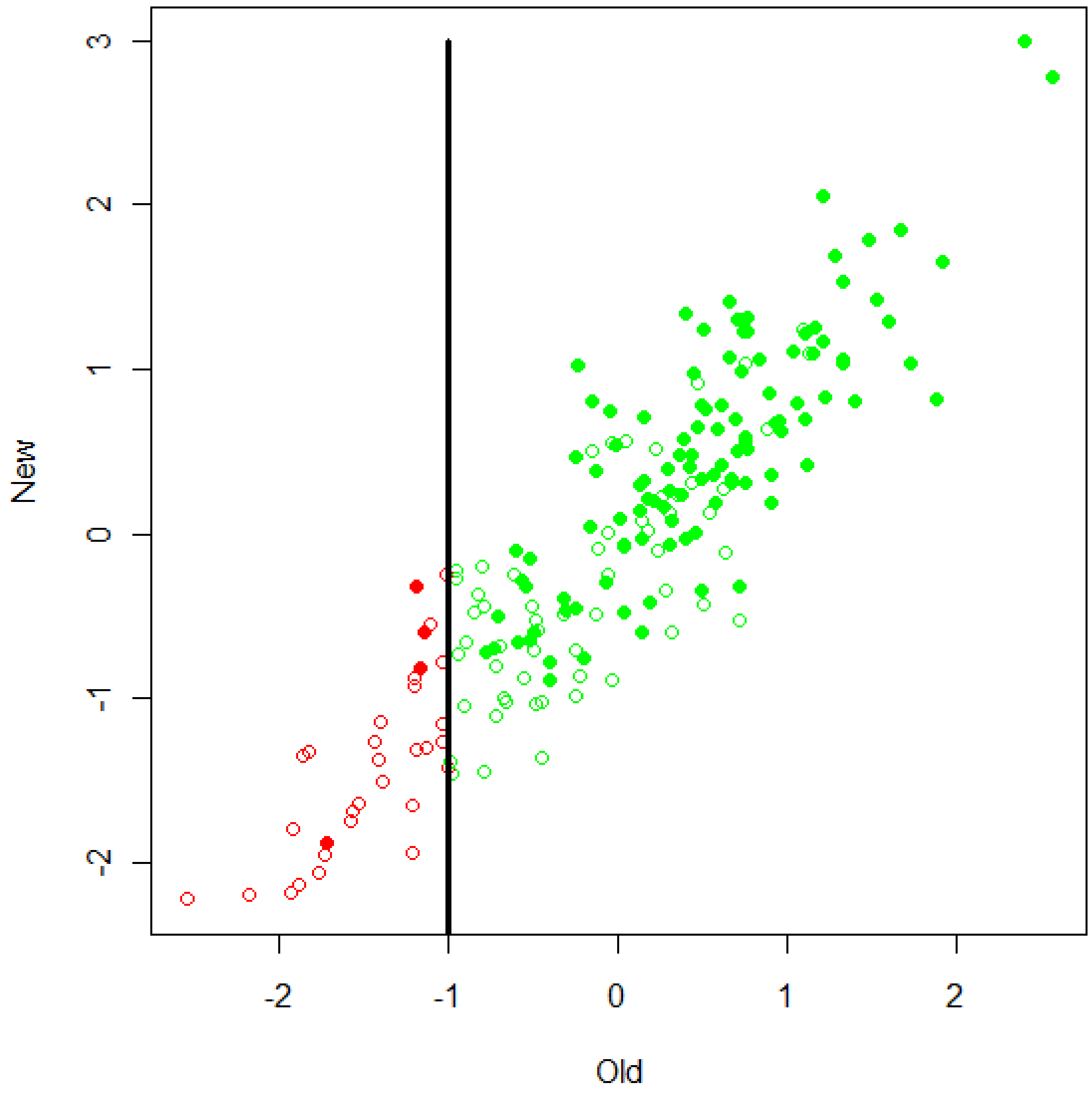

New
Old
3
2
1
0
-1
-2
-2
-1
0
1
2

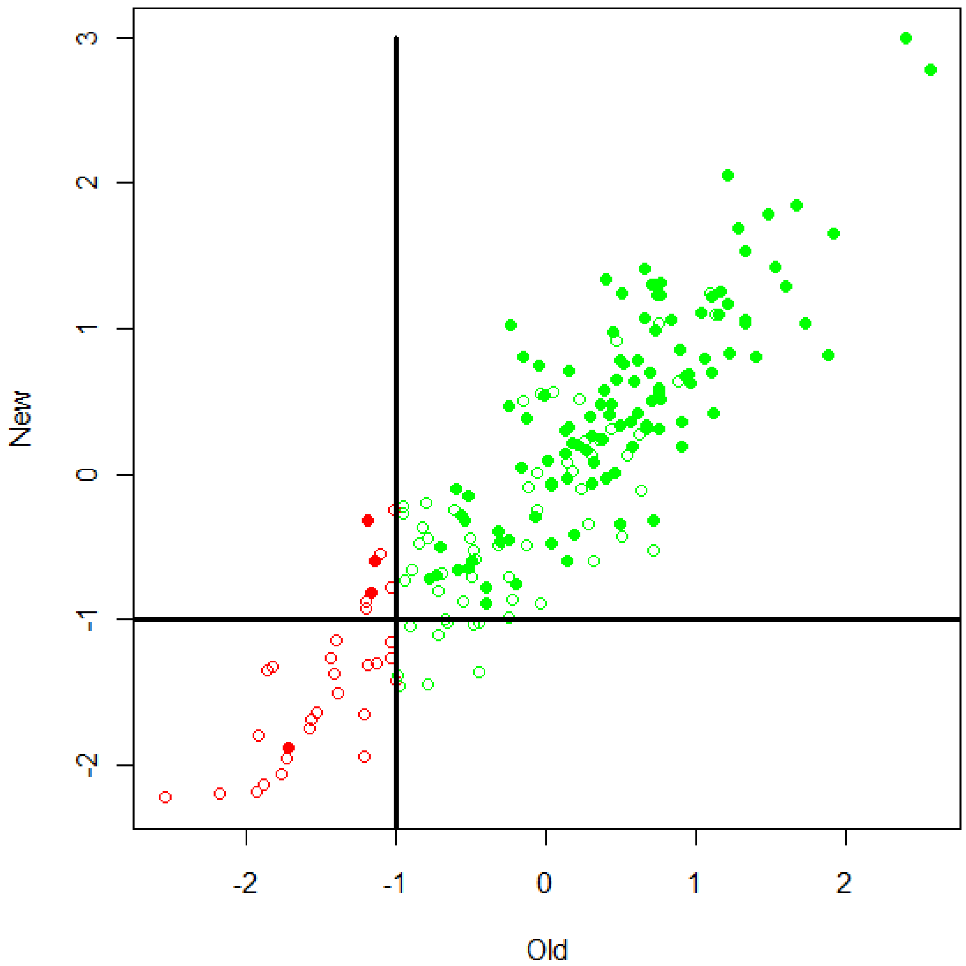
Bad class scores

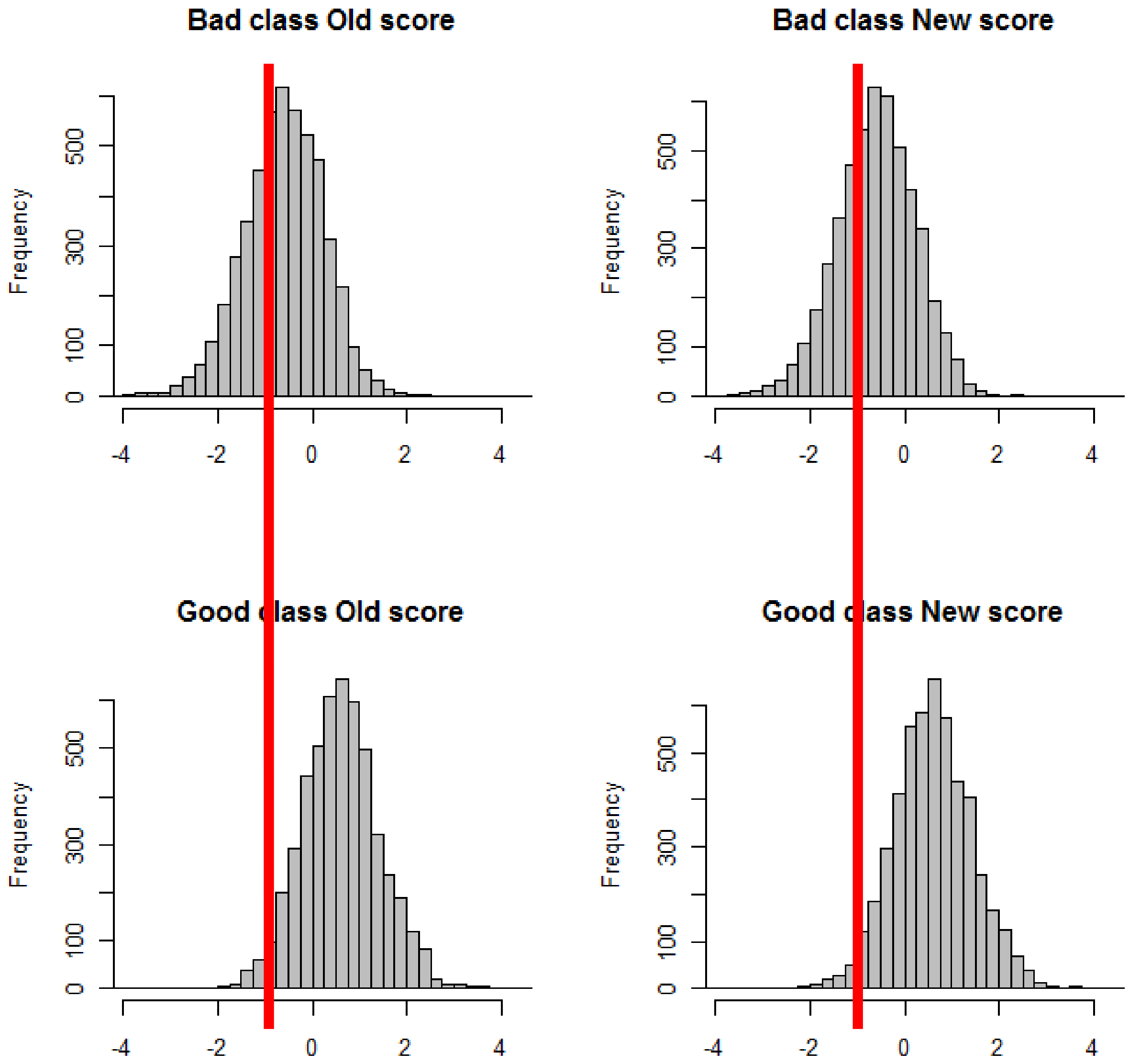


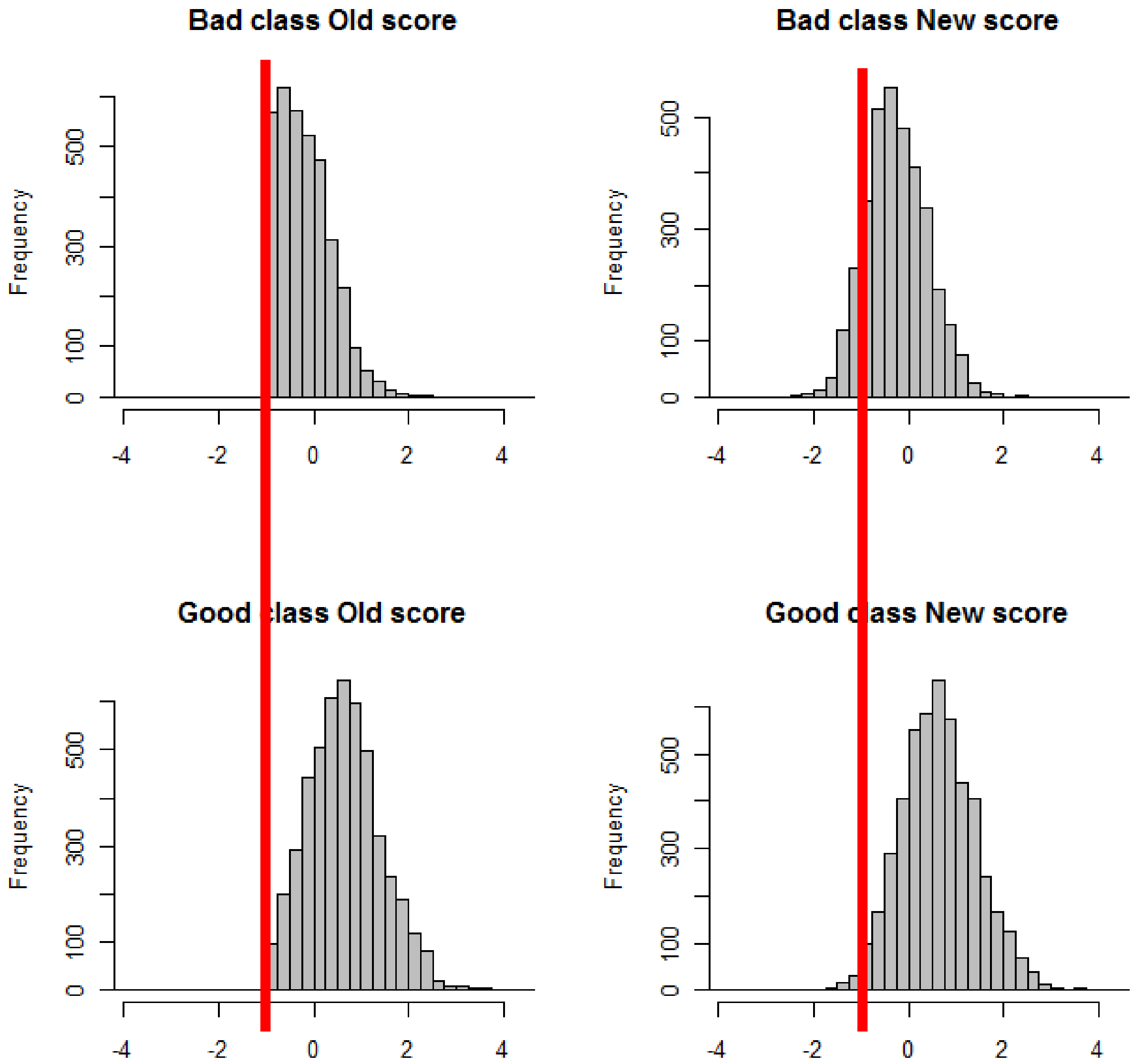
Good class scores




3
2
1
0
-1
-2
New
-2
-1
0
1
2
Old


New
3
2
1
0
-1
-2
-2
-1
0
1
2
Old


Bad class Old score
Bad class New score
Good class Old score
Good class New score
Frequency
0
100
300
500
-4
-2
0
2
4


Bad class Old score
Bad class New score
Good class Old score
Good class New score
Frequency
0
100
300
500
-4
-2
0
2
4

The effect:

Favours the new scorecard

leading to:

bias in favour of *TopScore's* new SVM model, even though it is really no better at all

unnecessarily incurring costs and risk of replacing old by new

Example 3: Fraud detection

Issue 1: What is a good system?

One which

'classifies fraudulent transactions as fraudulent, and legitimate transactions as legitimate' ?

But: no method is perfect
Need: criteria for assessing effectiveness

Timeliness: time scale: count of fraud transactions misclassified

Different weights on two kinds of misclassification

		True class	
		Fraud	Legitimate
Predicted class	Fraud	*a*	*b*
	Legitimate	*c*	*d*

A very well known consumer credit organisation evaluates fraud using the two ratios

$$R_1 = a/(a+c) \qquad \text{(\textit{recall, sensitivity})}$$

$$R_2 = b/(a+b) \qquad \text{(1- \textit{precision})}$$

In itself, this would appear to be fine

But in fact, the units of assessment used are ***accounts***

An account is flagged as potentially fraudulent if ***at least one transaction is so flagged***

Problem 1: This means that one can make the probability of flagging an account as fraudulent as near to 1 as one wishes by examining enough transactions

Problem 2: Fails to include *timeliness* in the measure

A superior measure

An ***epoch*** is a transaction sequence ending with either

(i) a *fraud flag* on a true fraud

Or

(ii) or end of observed sequence

n n n n f n n f n n **n** n n f n n n n n **n** n n n n n n **f**

		True class	
		Fraud	Legitimate
Predicted class	Fraud	a	b
	Legitimate	c	d

n n n n f n n f n n **n** n n f n n n n n **n** n n n n n n **f**

		True class	
Predicted class		Fraud	Legitimate
	Fraud	1	2
	Legitimate	3	21

This matrix includes ***timeliness*** in the count *c*

		True class	
		Fraud	Legitimate
Predicted class	Fraud	a	b
	Legitimate	c	d

$$cost = a + b + kc$$

worst case $cost = k(a+c) + b + d$

k = relative cost of 'misclassifying fraud as legitimate' versus converse

Overall performance measure for given threshold:

$$T_1 = (a+b+kc)/(k(a+c)+b+d)$$

This is very different from misclassification rate

$$e = (b+c)/(a+b+c+d)$$

Issue 2: Bias in evaluation

True transaction state sequence

nnnnnnnnnnnnnfnfnfffnfff

Detector D1 in place

Detector D2 proposed new detector

$D_i,\ i=1,2$ taking values 0 (no fraud suspected)

1 (fraud suspected)

$D_1 = 1$ and true state *n* means:

investigation and then sequence continues

$D_1 = 1$ and true state *f* means

investigation and then sequence ends

AND

true states of all previous transactions discovered

Define, for $j,k=0,1$

$$p_{jk}^{(n)} = P\left(D_1 = j, D_2 = k \mid n\right)$$

and

$$p_{jk}^{(f)} = P\left(D_1 = j, D_2 = k \mid f\right)$$

Then the new detector, D2, is unequivocally better than the old one, D1, if both

(i) $P\left(D_2 = 1 \mid f\right) > P\left(D_1 = 1 \mid f\right)$

and

(ii) $P\left(D_2 = 1 \mid n\right) \le P\left(D_1 = 1 \mid n\right)$

These are equivalent to

(i) $$p_{01}^{(f)} > p_{10}^{(f)}$$

and

(ii) $$p_{01}^{(n)} \leq p_{10}^{(n)}$$

		D2	
		0	1
D1	0	p_{00}	p_{01}
	1	p_{10}	p_{11}

Consider straightforward estimates of the $p_{jk}^{(f)}$ and $p_{jk}^{(n)}$ based on proportions of observations in

n		D2			***f***		D2	
		0	1				0	1
D1	0				D1	0		
	1					1		

BUT: All observed sequences in which a fraud is detected end in either the $\left(D_1 = 1, D_2 = 0\right)$ ***cell or the*** $\left(D_1 = 1, D_2 = 1\right)$ ***cell of the f table***

Consider a single terminating account with *c* fraudulent transactions

The (*c*-1) undetected frauds contribute only to f_{00} or f_{01}

Hence $E\left(f_{0k} \mid c \right) = (c-1)\, p_{0k}^{(f)} \big/ p_{0+}^{(f)}$

The one final detected fraud contributes to f_{1k}

Hence $E\left(f_{1k} \mid c \right) = p_{1k}^{(f)} \big/ p_{1+}^{(f)}$

So the expectations of simple multinomial estimates are

$$E\left(\tilde{p}_{0k}^{(f)} \mid c \right) = \left(1 - 1/c\right) p_{0k}^{(f)} \big/ p_{0+}^{(f)}$$

$$E\left(\tilde{p}_{1k}^{(f)} \mid c \right) = \left(1/c\right) p_{1k}^{(f)} \big/ p_{1+}^{(f)}$$

e.g. suppose $c = 1$

Then $$\tilde{p}_{0k}^{(f)} = 0$$
$$E\left(\tilde{p}_{1k}^{(f)} \mid 1\right) = p_{1k}^{(f)} / p_{1+}^{(f)} \geq p_{1k}^{(f)}$$

So the condition for D2 beating D1, that $p_{01}^{(f)} > p_{10}^{(f)}$ *cannot be met by these estimators*

If you use the simple multinomial estimators there is an intrinsic built-in bias favouring the existing detector

Example 4: Discrimination

Credit scoring is fundamentally *discriminatory*; seeks to discriminate good risks from bad risks

(c.f. discriminate good students from bad, safe drivers from unsafe, ...)

So: make the scorecard as effective as possible

The better the model, the more effective the bank, the lower the interest needed to cover the cost of defaulters

So: include all potential predictive variables we can think of

BUT:

US Equal Credit Opportunity Act, 1974:

it is illegal for creditors to discriminate against any applicant on the basis of race, colour, religion, national origin, sex, marital status, or age

Similar conditions in other countries

Even though (for example)

- women are generally less risky than men
- older men are generally less risky than younger

The Act makes it illegal to treat differently people who belong to certain groups with known different degrees of risk

So, as a consequence,

credit scorecards do not include sex as a predictor variable

to the disadvantage of the lower risk female class

who therefore have their loan applications rejected more often than their risk probability would justify

Solution:

If sex is a predictor, but the law says can't include sex, then include *a proxy* variable, Y, highly correlated with sex

Until the law catches up and makes Y illegal

Knocking out these variables risks knocking out further predictive power independent of sex

Solution depends on the real aim

Aim (A)*: 'treat men and women equally' in the sense that the same proportion of men and women are allocated to the good (and hence also bad) class*

Aim (B): *build the best risk classification model we can, but one which does not let sex contribute to our classification, even via variables we haven't thought of*

Illustrate with simple model: score is weighted sum of predictors, compare with threshold

(A)

'treat men and women equally' in the sense that the same proportion of men and women are allocated to the good (and hence also bad) class

Build separate models for men and women

Choose thresholds for each model so same proportion are accepted

'Fair' in the sense that equal proportions are accepted

'Unfair' in the sense that the risk thresholds differ

(B)

build the best risk classification model we can, but one which does not let sex contribute to our classification, even via variables we haven't thought of

Build a single model using all the variables we can think of, *including sex and any proxies*

then make classification using a score from the model but ignoring sex

'Fair' in the sense that it makes a decision in the predictor space orthogonal to sex

'Unfair' in the sense that the decisions are not based on the best possible estimates of default probability

Current situation is neither (A) nor (B)

So that

neither are the conditions of the ECOA being met

nor are people being assigned to a risk class on the basis of the best estimates of their default probability

The law has got itself into an ethical twist

Relevance:

Earlier this year the European Court of Justice ruled that it was unlawful sex discrimination for ***insurers*** to distinguish between men and women when deciding premiums

despite the fact that women are safer drivers and live longer

1978: female employees sued Los Angeles Department of Water and Power

on the grounds that they had to pay larger pension contributions, and thus took home less money (based on longer life expectancy)

Their case was successful

Issues:

- people should be treated as individuals, rather than stereotyped. But probability estimates must be based on groups, not individuals

- it is intrinsic to a *civilised* society that some risks subsidise others. No fault risks (e.g. genetic) might be subsidised. What about lifestyle risks (e.g. smoking)? Moral hazard

- does our performance criterion refer to individuals or populations?

Suppose cost of driving insurance is equalised at a weighted mean of men and women

Then more of the high risk category will be encouraged to drive (as it’s cheaper to take out insurance), and fewer of the lower risks will be encouraged to drive (as it’s more expensive)

But paying for insurance does not change the risk

So that the risk to all of us is increased

Conclusions:

It’s not enough to develop an effective algorithm

Each problem is different

It’s crucial to match the solution to the problem

The details count

thank you !