Selection Bias in Credit Scorecard Evaluation

David J. Hand and Niall M. Adams

Department of Mathematics
Imperial College London

August 2011

Contents

Selection Bias

Perennial problem arising from making inferential statements about a population based on a non-random sample from the population.

Familiar example: reject inference $\rightarrow$ problem of constructing a scorecard from the entire population, when available data only refers to previously accepted customers.

The scorecard may not perform well on the whole population, due to these "holes" in the data space.

Reject inference methods attempt to infer the good/bad status of previously rejected applicants $\rightarrow$ that is, to fill in the data space.

Can't get something for nothing, so RI methods attempt to find a source for this missing information. Sadly, this will be the case for this talk also!

Scorecard Evaluation

This talk is concerned with the impact of selection bias on comparative application scorecard evaluation.

Scorecard evaluation is essential to the industry. Performance deteriorates, related to changes in

- applicant population
- economic climate
- competitive environment

Scorecards are updated regularly. Such updating involves calibrating the performance of a new scorecard against an existing scorecard, to determine whether it is superior.

Moreover, each stage of the scorecard construction process (such as determining whether a new characteristic improves performance) can be regarded as comparative performance evaluation

Framework

Suppose we have an application scorecard (the old scorecard) and we wish to see how another (the new scorecard) performs relative to the old.

Apply both scorecards to the same sample of customers. These customers have been accepted by the old scorecard, and this produces an asymmetry. This produces an unfortunate effect on popular performance metrics.

Punchline: Selection bias in this framework can lead to incorrectly favouring a new scorecard over an existing scorecard, and raises the costs and risks in unnecessarily replacing the new scorecard with the old.

Demonstration and Simulations

Notation:

- $f(s_o, s_n)$ joint density of old and new scores for bad class
- $g(s_o, s_n)$ joint density of old and new scores for good class
- f_o, f_n, g_o, g_n corresponding marginal densities
- F_o, F_n, G_o, G_n corresponding marginal cumulative distribution functions

Assume (stochastic dominance):

- $F_o(s) > G_o(s)$
- $F_n(s) > G_n(s)$

Not restrictive, implies ROC curve never crosses *chance* diagonal.

Application scorecards select customers by comparing score s with threshold t

$$\text{decision} = \begin{cases} \text{accept} & s > t \\ \text{reject} & \text{otherwise} \end{cases}$$

Selection process based on old scorecard, so data in test sample based only on applicant with score pairs

$$\{(s_o, s_n)\} : s_o > t$$

Broadly, truncation of the old scorecard is absolute, while the truncation of the new scorecard depends on the correlation between old and new scorecards.

Precise impact of this truncation, and size of resulting bias, will depend on the shape of the various score distributions.

To illustrate the effect, let old and new scorecard joint distributions be bivariate normal, yielding identical marginal score distributions for both good and bad classes, except for a difference in the mean.

This means that the old and new scorecards produce identical performance measures when applied to the *entire* population: neither is better than the other.

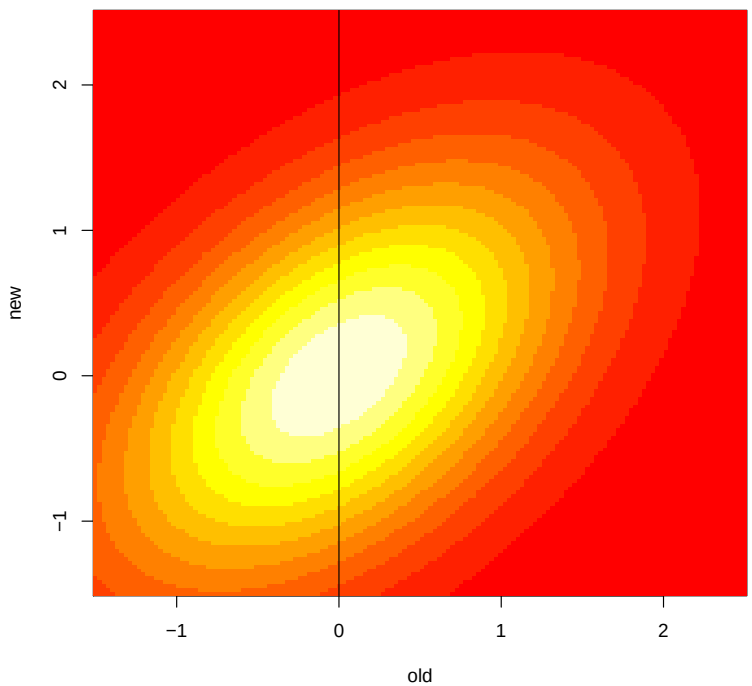
For simplicity, assume mean of new and old bad score distribution is zero, and common mean of good score distribution is $\mu > 0$. Similarly, common standard deviation of all distribution is $\sigma = 1$.

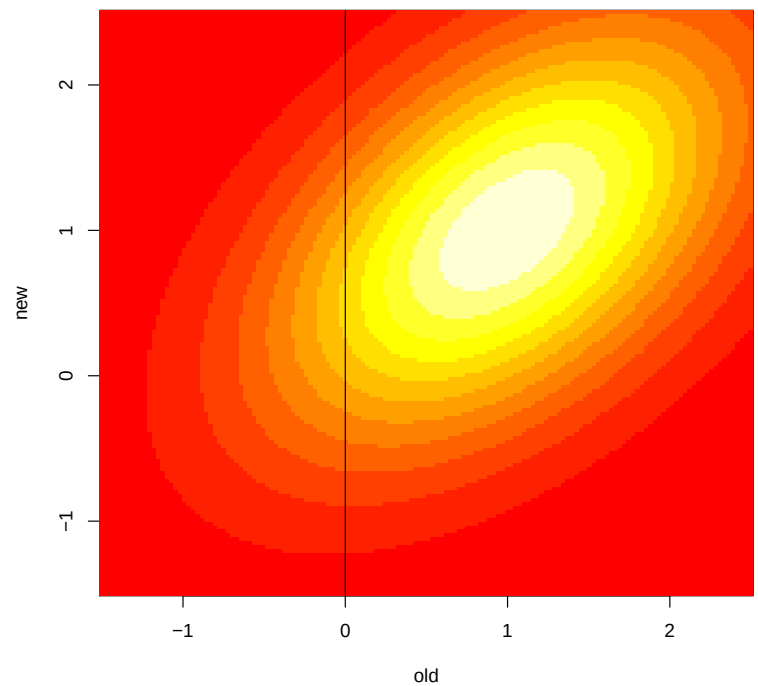
Thus,

$$f(s_o, s_n) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(s_o^2 + s_n^2 - 2\rho s_o s_n)\right\}$$

where ρ is the correlation between the old and new scores, and

$$g(s_o, s_n) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[(s_o^2 - \mu) + (s_n^2 - \mu) - 2\rho(s_o - \mu)(s_n - \mu)\right]\right\}$$





Old Scorecard

For the old scorecard, when the distribution of scores is truncated at t, so only those with $s > t$ are retained, the marginal density for bad cases is

$$f_{ot}(s) = \frac{\phi(s)}{(1 - \Phi(t))} = \frac{\phi(s)}{\Phi(-t)}, \quad s > t$$

where

- ϕ is the standard normal PDF
- Φ is the standard normal CDF

Similarly, the marginal (truncated) score density for good cases

$$g_{ot}(s) = \frac{\phi(s - \mu)}{(1 - \Phi(t - \mu))} = \frac{\phi(s - \mu)}{\Phi(\mu - t)}, \quad s > t$$

New Scorecard

Marginal distribution of the new scorecard for bads is

$$\begin{aligned} f_{nt}(s) &= \frac{\phi(s)\left\{1-\Phi(t-\rho s)/\sqrt{1-\rho^2}\right\}}{1-\Phi(t)} \\ &= \frac{\phi(s)\Phi\left((\rho s-t)/\sqrt{1-\rho^2}\right)}{\Phi(-t)} \end{aligned}$$

and for the goods

$$\begin{aligned} g_{nt}(s) &= \frac{\phi(s-\mu)\left\{1-\Phi(t-\mu-\rho(s-\mu))/\sqrt{1-\rho^2}\right\}}{1-\Phi(t-\mu)} \\ &= \frac{\phi(s-\mu)\Phi(\mu+\rho(s-\mu)-t)/\sqrt{1-\rho^2}}{\Phi(\mu-t)} \end{aligned}$$

Performance measures

The densities above are have corresponding CDFs denoted F_{nt}, F_{ot}, G_{ot}, G_{nt}.

Having this simple framework provides the chance to reason about the extent of overlap of the distributions, noting that there should be none.

Three standard measures, given distribution functions F and G

- KS statistic: $max_s|F(s) - G(s)|$ (see DJH talk on Friday)
- AUC: $\int F(s)dG(s)$
- H measure: (a coherent performance measure, the same for all scorecards. See Hand (2009) for details).

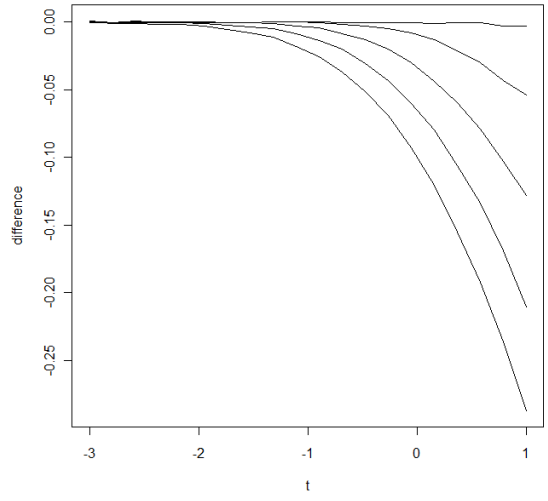
Since concerned with difference between score distributions, define

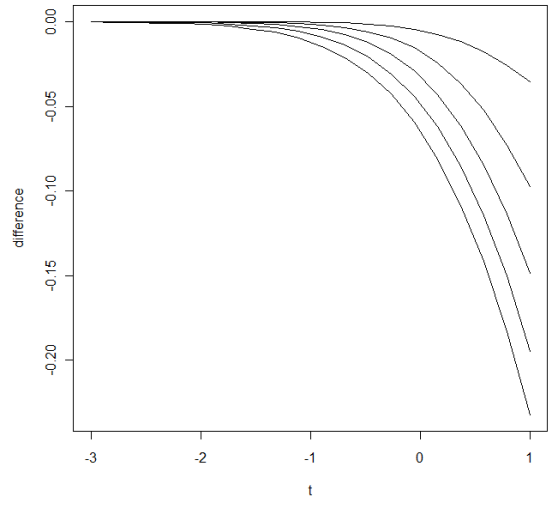
- $P_{KS} = \max_s |F_{ot}(s) - G_{ot}(s)| - \max_s |F_{nt}(s) - G_{nt}(s)|$
- $P_{AUC} = \int F_{ot}(s) dG_{ot}(s) - \int F_{nt}(s) dG_{nt}(s)$
- P_H difference in H measure

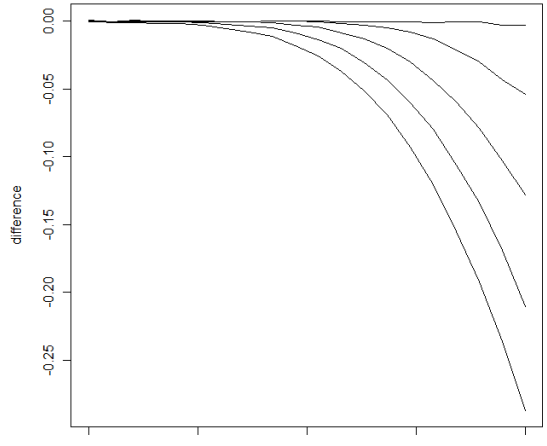
In all cases, negative values imply that the new scorecard is better than the old scorecard.

Simulation results I

Here, $\mu = 2$ and t varies. Each plot displays performance measure, with decreasing lines corresponding to decreasing correlation. (Left - KS, right - AUC, bottom - H).







In all cases, as t increases, the preference for the new scorecard increases. Recall that this is only an artifact of the biased sample - with the full population, all distributions are the same.

As expected, the effect becomes more marked with decreasing correlation between old and new scores.

Simulation results II

Interesting to consider performance with μ. Following tables refer to AUC, but other measures exhibit similar behaviour.

Note that the choices of μ are selected to explore the range of performance. For example

μ	Population AUC
1.9	0.67
2.2	0.76
2.5	0.83
2.8	0.88

Perhaps $\mu = 2.2$ is most representative of typical application scorecards?

AUC results

$\rho = 0.2$,

t	$\mu = 1.9$	$\mu = 2.2$	$\mu = 2.5$	$\mu = 2.8$
0.1	-0.112	-0.073	-0.033	-0.021
0.5	-0.185	-0.129	-0.062	-0.039
0.9	-0.280	-0.209	-0.110	-0.073

$\rho = 0.8$,

t	$\mu = 1.9$	$\mu = 2.2$	$\mu = 2.5$	$\mu = 2.8$
0.1	-0.003	-0.016	-0.005	-0.003
0.5	-0.005	-0.034	-0.014	-0.008
0.9	-0.008	-0.064	-0.032	-0.020

The effect is

- again stronger for low correlation than high correlation.
- more profound the less separable are the distributions (this is expected, as the problem gets easier, the difference will become smaller).

If we consider $\mu = 2.2$, then the effect might be strong enough, even for high correlation, to encourage us to change scorecard?

In that case, we have been misled, simply by a selection effect. Replacing the scorecard incurs costs for software and staff retraining. Perhaps better to err on the side of caution?

Real Data Example

- UPL data from major UK bank (1994-1997).
- 42K cases, on 30 variables
- data pre-scored and subject to previous variable selection
- clear evidence of population drift (examples in Adams et al. (2010)

Population drift is a potential confounder with this bias. To remove the effect of drift, data order randomised.

Note, that this is simply meant to be an illustration of the bias demonstrated above, present in real data.

Old model (LR), built on 20K observations. 6 variables, including

- timing indicator
- loan amount
- search indicator
- credit card indicators

Now, consider 22K observations of test data.

Select the 80% best scores from the old model. This set of applicants constitutes the test set, for comparing new model with old.

New model (LR), built on same data. 15 variables, including

- some old model variables
- age indicators
- address indicators
- employment indicators

Correlation of scores: 0.37

- selected sample: AUC(old)=0.633, AUC(new)=0.661
- full test data: AUC(old)=0.681, AUC(new)=0.676

Again, this is not to be a realistic scorecard construction exercise. The effect of the bias here is small - but still misleading.

Conclusion

- Have demonstrated another problem arising from selection bias - a preference for a new scorecard in comparative evaluations.
- Like other selection bias problems, any solution would require extra information. Solutions unlikely to be procedurally palatable.
- Interesting interaction between this problem, and population drift. What other sources contribute to complicate these comparisons?

References

Hand, D.J. (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, **77**, 103-123.

Hand, D.J. and Adams, N.M. (2011) Selection bias in credit scorecard evaluation. Technical Report, Department of Mathematics, Imperial College London.

Adams, N.M., Tasoulis, D.K., Anagnostopoulos, C. and Hand, D.J. (2010) Temporally-adaptive linear classification for handling population drift in credit scoring. In *COMPSTAT2010*, Proceedings of the 19th International Conference on Computational Statistics, Lechevallier, Y. And Saporta. (Eds), 2010, Springer, 167-176.