Should we throw away the variance of the score estimates?

Mark Surdutovich (Itaú Unibanco)

&

Genaro Dueire Lins (Itaú Unibanco)

Should we throw away the variance of the estimate?

- The standard practice in credit scoring is to run a logistic regression and discard the predicted values and standard errors of each score

- We argue that this can lead to discarding of meaningful information and possibly of a way to control and price model risk

- We show that punishing the scores with biggest predicted variances can produce modest but “free” gains in score performance especially in smaller portofolios like SMEs

- We also show that, if the score volatility is correlated with macroeconomic conditions, it can make the portofolio worth 0,5% less then it apparently does.

What is wrong with throwing away the variance of the estimates:

- The logistic regression produces as output: $X * \beta$ and the estimate of variance of every estimate

- The $X * \beta$ is used to order clients. It is assumed that it contains ALL the ordinal information from the logistic regression

- This is not generally the case...

- If x is a random variable, $E(f(x)) \neq f(E(x))$ for a non-linear function $f(x)$

- In particular, if $E(x)>E(y)$ it is not necessarily the case that $E(f(x))> E(f(y))$

- In our case, $X * \beta$ passes though a non-linear transformation to become what we actually want to order clients on:

$$\text{Probability of default} = \frac{e^{X_{\mathrm{i}} * \beta}}{1+e^{X_i * \beta}}$$

Punishing volatile scores

- The probability function is strictly convex from $\mathrm{X_i} * \beta < 0$ (corresponding to probabilities of default less then 0.5) and strictly concave if $\mathrm{X_i} * \beta > 0$

- If the are two estimates $\mathrm{X_i} * \beta = \mathrm{X_j} * \beta < 0$ but with $\sigma_j = 0$ and $\sigma_i > 0$ by Jensen´s inequality we obtain:

$$P\ of\ default_i > P\ of\ default_j$$

<u>The standard procedure in credit scoring, however, would have assigned $\mathrm{X_i} * \beta$ and $\mathrm{X_j} * \beta$ the same score</u>

Luckily, there is a handy approximation

- Crooks 2007 derived a simple approximation for the probability given $\mathrm{X_i} * \beta$ and σ_i:

$$Probababily \approx 1/(1 + e^{-\frac{X_i \beta}{\sqrt{1+\frac{\pi}{8}\sigma_i}}})$$

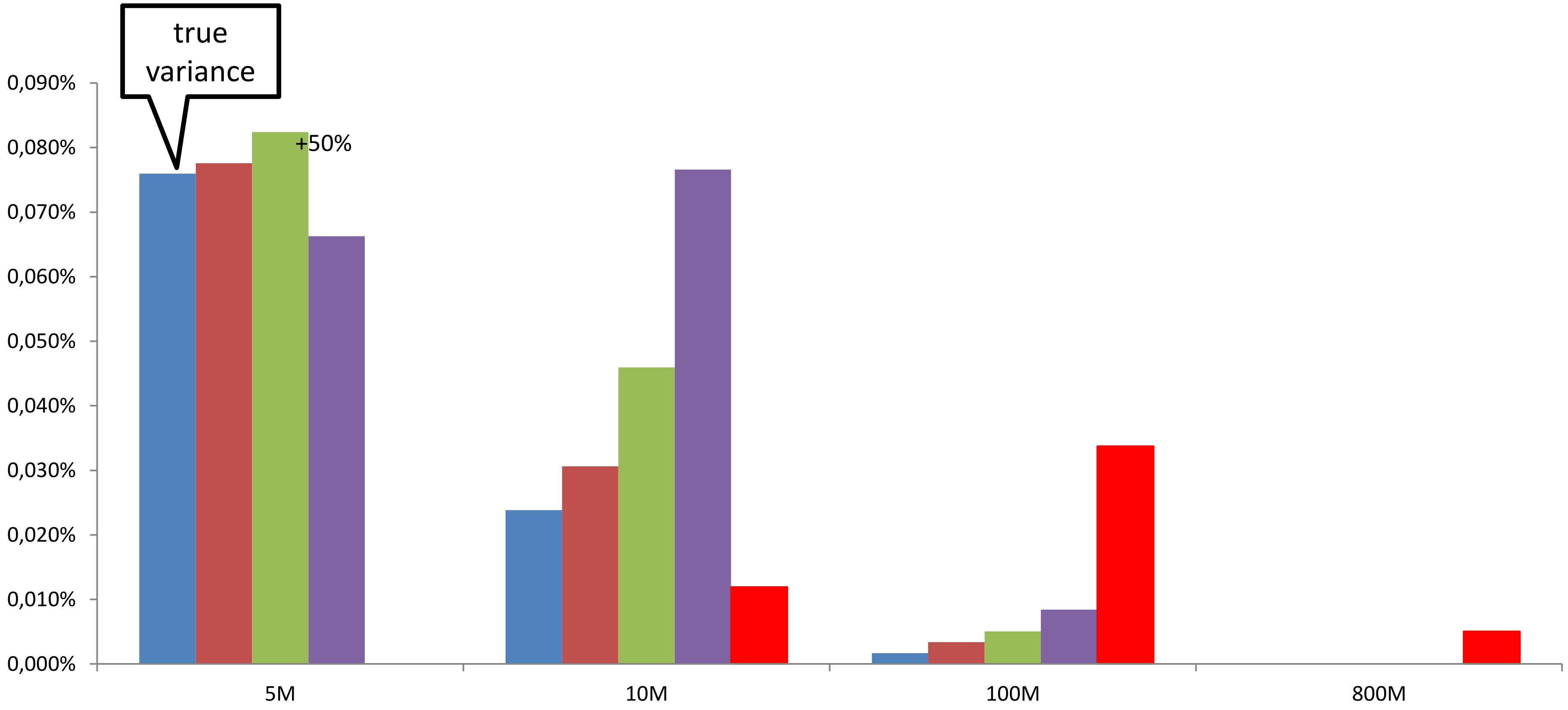
Small gains in small samples:

- **Model 1**: reestimation of the equation of market performance monitoring model on a sample of 5 thousand showed a gain of relative Gini around 0.08%, this gain declined to 0,02% for a sample of 10 thousand and disappeared on larger samples

- **Model 2**: behavioral model for SMEs with a sample of over 25 thousand estimated on a single vintage. The adjustment showed no gains

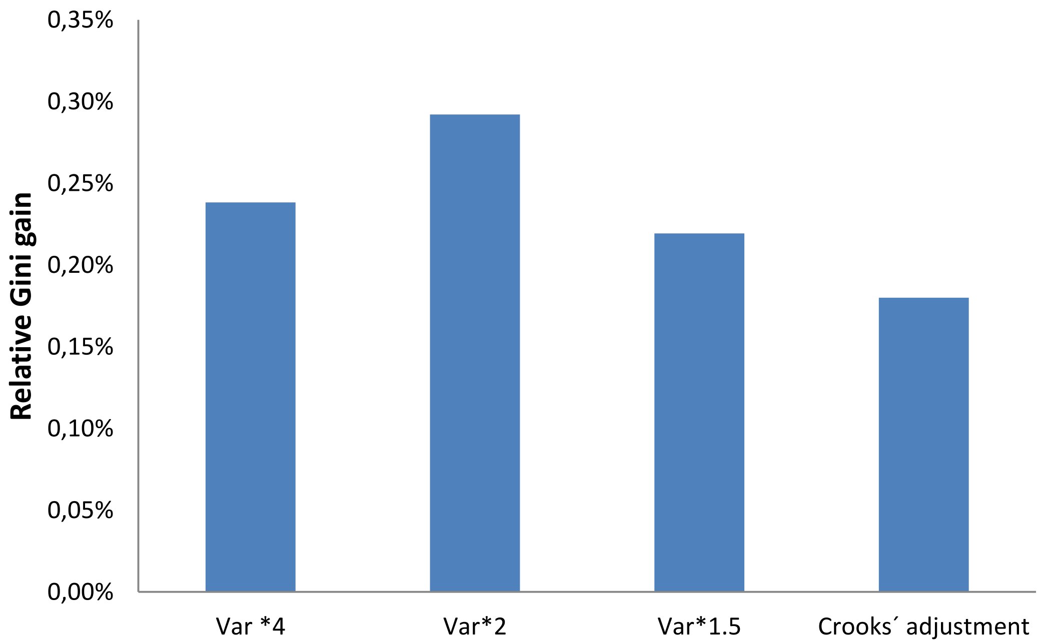
However, we found something interesting...

Boosting the variance in the Crooks formula produced gain in both models, increasing in sample size:



...and the results for very small samples show larger gains

- Model 1 was modified with just 3 variables ramaining and applied estimated to 15 samples of 700 clients
- The gains of relative Gini are more sizable with variance boosting still giving superior results:



If it´s plausible that the variance of the score is capturing the effect of economic conditions...

- …because depending of more volatile variable would produce a volatile score.

- How large can this effect be? To find this bound, we assumed that all the variance of the estimates for Model 2 was related to market risk and applied the CAPM model with parameters from Estada 2003 (the risk premium being 5.5% and risk free rate 5%).

- Assuming loan of same size, the CAPM price of portfolio was 0.5% lower that if variance of the estimates was assumed uncorrelated to market risk. For comparison, increasing all the expected probabilities of default by 50% led to a decrease of CAPM price by 11%

- More interestingly, the relative CAPM prices of each score could be used a new ordering, potentially more robust to macroeconomic shocks

Next steps

- Both models tested were behaviour models. We are hoping to test the method on an application model with continuous variances, were we expect to obtain bigger gains

- We want to test the hipothesys that volatile scores are more sensitive on marcoeconomic conditions