During our free webinar earlier this month, "Actual vs Expected: Statistical Framework for Scorecard Management" we invited participants to send us questions. I am posting this one on the use of Marginal Information and the process ordering for selecting variables to enter a scorecard during the model build.

Question

"How would you suggest processing the coarse bin?

With many variables available it is difficult to coarse bin each one individually. The initial indication comes from Information Value (IV) then Marginal Information Value (MIV). This helps determine the variables to select, but that selection would depend on the fine bins.

If you select a fine binned variable from IV or MIV then re-bin, would this change its IV/MIV?

What is the correct/suggested order – should you use the coarse bin before or after? If before, how do you select what to coarse bin?

Or is there some sort of adjustment/convergence monitoring that can be done with the MIV each time a variable is re-binned?"

Answer from Gerard Scallan, ScorePlus consultant.

1. MIV is not very sensitive to the classing used. My recommendation is to run the MIV analysis on fine classed characteristics - in other words, you only look at classing when you decide to bring a certain predictor into the model.

   This means that you over-estimate the "real" MIV because there is some noise from small samples. However, this is not a problem: if the estimated MIV is "small" then the real MIV is even smaller. Also, for anything that comes into the scorecard, you can then do a classing and see a less noisy MIV.

   Incidentally, this is why I focus on the MIV rather than the Marginal Chi^2 - the p-level of the Chi^2 depends on the number of degrees of freedom which is much too high before the coarse classing.

2. An even better solution is first to select an extra characteristic to add to the model. You run a stepwise logistic regression with dummy variables corresponding to partitions at each fine break in all the selected characteristics - including the new one. Then the stepwise process will do the classing for you, by only bringing in the fine breaks which are significantly different from their neighbours. It also means that each prior characteristic is re-classed when a new one is added to take account of interactions.

   This means that:
   a) you don't class anything that doesn't make it into the final model and
   b) for those that are used in the model, you can (largely) automate the classing process.

   I say "largely" because you still need to sense-check the predictors and classing from the process. This is described in slides 718-722 of the Scorecard Building Strategies in the Building Better Scorecards course. It's the ultimate approach to classing.

3. You can also avoid the problem by using the Marginal Kolmogorov-Smirnov measure in place of MIV. You can find details in **www.scoreplus.com/assets/files/Marginal-KS-analysis-Measuring-lack-of-fit-in-logistic-regression-Edinburgh-conference-Aug-2013.pdf**. This just uses rank-ordering and doesn't bother with classing.