

Evidence on the Value of Data Mining: Reporting Back

ELMAR Contribution by J. Scott Armstrong
May 21, 2003

I received only four replies to my ELMAR request (March 24, 2003) for evidence on the value of data mining. Where are all those advocates of data mining? One person said that he had long been preaching that data mining is useless. One person thought that I should do more reading on the topic, and another provided things for me to think about. No one provided a source with empirical evidence showing that data mining can improve decision-making or forecasting, but a reply from Paul Bottomley provided negative evidence. He and Agnes Nairns designed an experiment to see whether managers could distinguish between cluster analysis outputs derived from real and random data. Random data results were perceived as equally useful as real data results for purposes of market segmentation. Their paper is titled "Blinded by Science: The Managerial Consequences of Inadequately Validated Cluster Analysis Solutions." Paul's e-mail is BottomleyPA@Cardiff.ac.uk

Their study reminded me of my attempts to eliminate earlier versions of data-mining approaches in my paper, "Tom Swift and his Electric Factor Analysis Machine, and in Tom's further adventures as a researcher for the International Caribou Chip Co., where he used step-wise regression analysis (see "How to Avoid Exploratory Research"). These papers are in full-text under "applied statistics" at <http://www.jscottarmstrong.com>

My earlier hypothesis was too conservative: It stated that data-mining was useless. I think a directional hypothesis is in order: *Data mining is harmful because it can be expensive and, more importantly, it deceives people into thinking that they have gained useful insights.*