White Paper

It is a Mistake to… Accept Leaks from the Future

by John Elder

Founder & CEO

August 2014

ELDER RESEARCH

DATA SCIENCE & PREDICTIVE ANALYTICS

This article is Part 6 (of 11) of a series by the author on the Top 10 Data Mining Mistakes, drawn from the *Handbook of Statistical Analysis and Data Mining Applications*.

I often evaluate promising investment systems, for possible implementation. In one, a Ph.D. consultant, with a couple of books under his belt, had prepared a neural network model for a Chicago bank to forecast interest rate changes. The model was 95% accurate – astonishing given the importance of such rates for much of the economy. The bank board was cautiously ecstatic, and sought a second opinion. My colleagues found that a version of the output variable had accidentally been made a candidate input. Thus, the output could be thought of as only losing 5% of its information as it traversed the network.

One investment system we were called in to examine was 70% accurate in forecasting the direction a market index would move the next day. Its developers were quite secretive, but after a great deal of work on behalf of the client considering investing, I eventually duplicated its actions exactly with a simple moving average of three days of prices. This simplicity was disappointing, but much worse was that the three days were centered on *today*. That is, tomorrow's price was one of the inputs! (They'd have had 100% accuracy, if they'd just dropped one of the input variables.) Another trading system, developed with monumental effort over several years and involving the latest research in Genetic Algorithms (GA), focused on commodities. Eventually, I 99.9% matched it with, essentially, two lines of code, which made obvious that its use was impractical. That is, the complex GA devolved to a simple model (the flaws of which then became quite clear), in a manner impossible to discern by examining the extremely complex modeling machinery. In all these cases, the model's author was the chief one deceived.

One trick is to look hardest at any input variable that works too well. For instance, on a cross-sell project – trying to identify clients of an auto club who'd be good prospects for a more profitable insurance product – we found a code which was present about 25% of the time, but was always associated with insurance purchasers. After extended inquiry (as the meaning of data fields are often lost to the mists of time) we found that the code was the type of insurance cancellation; that is, that it really represented the fact that about a quarter of purchasers cancelled their insurance each year. Dorian Pyle, author of a thorough book on *Data Preparation for Data Mining*, has recounted privately that he's encountered problems that required seven such "decapitation" passes, where the best variable turns out to be a leak from the future.

In general, Data Warehouses are built to hold the best information known to date on each customer; they are not naturally able to pull out what was known at the time that you wish to study. So, when storing data for future mining, it's important to date-stamp records and to archive the full collection at regular intervals. Otherwise, recreating realistic information states will be extremely difficult, and will lead to wrong conclusions. For instance, imagine one wished to study whether "dot-com" companies were, in aggregate, really a bad bet.

Using a price-quoting service, one pulls down all the histories of current such companies and studies their returns. Quite likely, they would have been a great bet, despite the horrible shakeout in that market sector that started in roughly March 2000. Why? Were their early gains so great as to absorb later massive losses? Actually, one would have made a study error — "survivor bias" – by looking back from *current* companies, which is all most data services carry. A recreation of the set of companies that existed at the earlier time, including the doomed ones, would provide a much more realistic (i.e., negative) result.

About the Author

Dr. John Elder, Founder and CEO of Elder Research, leads the largest and most experienced data science consulting firm in the U.S. For 20 years, the team has applied advanced analytics to achieve high ROI for investment, commercial and security clients in fields from text mining and stock selection, to credit scoring and fraud detection. John has Engineering degrees from Rice and the University of Virginia, where he's an adjunct professor. He's authored innovative tools, is a popular keynote speaker, and has chaired International Analytics conferences. Dr. Elder served 5 years on a panel appointed by President Bush to guide technology for National Security. He has co-authored three books (on data mining, ensemble modeling, and text mining), two of which won Prose "book of the year" awards.

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

National Capital Region
2101 Wilson Boulevard
Suite 900
Arlington, VA 22201

855.973.7673

Headquarters
300 W. Main Street
Suite 301
Charlottesville, VA 22903

434.973.7673

Maryland Office
839 Elkridge Landing
Suite 215
Linthicum, MD 21090

855.973.7673