White Paper

It is a Mistake to…Lack Relevant Data

by John Elder

Founder & CEO

June 2013

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

This article is Part 1 (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*.

Mining data to extract useful and enduring patterns remains a skill arguably more art than science. Pressure enhances the appeal of early apparent results, but it is all too easy to fool oneself. How can one resist the siren songs of the data and maintain an analysis discipline that will lead to robust results? It is essential to *not*: lack (proper) data, focus on training, rely on one technique, ask the wrong question, listen (only) to the data, accept leaks from the future, discount pesky cases, extrapolate (practically and theoretically), answer every inquiry, sample casually, or believe the best model.

Introduction

It has been said that good judgment comes through experience, but experience seems to come through bad judgment! In two decades of mining data from diverse fields we have made many mistakes, which may yet lead to wisdom. Below, we briefly describe, and illustrate from examples, what we believe are the "Top 10" mistakes of Data Mining, in terms of frequency and seriousness. Most are basic, though a few are subtle. All have, when undetected, left analysts worse off than if they'd never looked at their data.

After compiling the list, we realized that an even more basic problem – mining without (proper) data – must be addressed as well. So, numbering like a computer scientist (with an overflow problem), here are the first of mistakes 0 to 10. [1]

Lack Data

To really make advances with an analysis, one must have labeled cases; i.e., an output variable. With input variables only, all one can do is look for subsets with similar characteristics (cluster), or find the dimensions which best capture the data variation (principle components). These unsupervised techniques are much less useful than a good (supervised) prediction or classification model. Even with an output variable though, the most interesting class or type of observation is usually the rarest by orders of magnitude. For instance, roughly 1/10 of "risky" individuals given credit will default within two years, 1/100 people mailed a catalog will respond with a purchase, and perhaps 1/10,000 banking transactions of a certain size require auditing. The less probable the interesting events, the more data it takes to obtain enough to generalize a model to unseen cases. Some projects probably should not proceed until enough critical data is gathered to make them worthwhile.

For example, on a project to discover fraud in government contracting, known fraud cases were so rare that strenuous effort could initially only reduce the size of the haystack in which the needles were hiding[2]. That is, modeling served to assure that the great majority of contracts were almost surely not fraudulent, which did enable auditors to focus their effort. But more known fraud cases — good for data miners, but bad for taxpayers — could have provided the modeling traction needed to automatically flag suspicious new cases much sooner. This was certainly the situation on another project, which sought to discover collusion on tax fraud. Unfortunately (for honest taxpayers), there were plenty of training examples, but their presence did lead to stronger, immediate modeling results, which was ultimately beneficial to taxpayers.

One can't mine without data, but not just any data will work. Many data mining projects have to make do with "found" data, not the results of an experiment designed to illuminate the question studied. It's like making a salad out of weeds found in the yard.

One sophisticated credit-issuing company realized this when seeking to determine if there was a market for their products in the class of applicants previously routinely dismissed as being too risky. Perhaps a low-limit card would be profitable, and even help a deserving subset of applicants pull themselves up in their credit rating? [3] But, the company had no data on such applicants by which to distinguish the truly risky from those worth a try; their traditional filters excluded such individuals from even initial consideration. So, they essentially gave (small amounts of) credit almost randomly to thousands of risky applicants and monitored their repayments for two years. Then, they built models to forecast defaulters (those late on payments by 90+ days) trained only on initial application information. This large investment in *creating* relevant data paid off in allowing them to rationally expand their customer base.

So, make sure that the data you're mining is relevant to the problem to be solved!

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.

[1] Most examples are from our own, and our colleagues' experiences, but some identifying details are mercifully withheld.

[2] Virtually all known cases were government workers who had, out of guilt, turned themselves in. Most, it seems, meant to pay back what they had fraudulently obtained (but how?). One audacious fraudster was discovered however, after coworkers realized that the clerk had been driving a different sports car to work every day of the month!

[3] For a decade now, the credit industry has mailed over a billion offers a year to American households; the high-risk market was one of the few places not saturated a few years ago. Credit profits are nonlinear with risk, and remind us of the triage system established during the Napoleonic wars, when the *levee en masse* swelled the battlefields and, combined with the new technology of cannons, etc., led to an army's medical resources being completely overwhelmed. Battlefield wounds were classified into three levels – the most minor to be passed by and treated later (if at all), more serious to receive immediate attention, but the most serious were judged not likely to be worth a physician's time. (We can envision a combatant, aware of hovering between the latter two classes insisting, like the Black Knight in the *Monty Python* movie, "What? The leg gone? It's just a flesh wound!") Likewise, credit companies make the most profit on individuals in the middle category of "woundedness" – those who can't pay off their balance, but keep trying. But they lose 5-10 times as much on clients just a little worse off, who eventually give up trying altogether. So, for models to be profitable at this edge of the return cliff they have to forecast very fine distinctions. Recent downturns in the economy have severely punished the stocks of companies that aggressively sought that customer niche – especially if they did not give obsessive attention to model quality.

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