White Paper

It is a Mistake to…Discount Pesky Cases

by John Elder

Founder & CEO

November 2014

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

This article is Part 7 (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*.

Outliers and leverage points can greatly affect summary results and cloud general trends. Yet, one must not routinely dismiss them; they could *be* the result. The statistician John Aitchison recalled how a spike in radiation levels over the Antarctic was thrown out for years, as an assumed error in measurement, when in fact it revealed a hole in the Ozone layer that proved to be an impressive finding. To the degree possible, visualize your data to help decide whether outliers are mistakes to be purged or findings to be explored.

I find the most exciting phrase in research not to be a triumphal (and rare) "Aha!" of discovery, but the muttering, "That's odd…" of puzzlement. To be surprised though, we must have expectations. So, I urge colleagues to make hypotheses of how results will turn out from their upcoming experiments. After the fact, virtually everything can and will be plausibly interpreted. A colleague of mine at ERI – Stein Kretsinger– shared an office with me when he was a Masters engineering student at the University of Virginia decades ago. He was working with medical data (often an extremely tough domain) and presented some interim findings as a graph on an unlabeled transparency to the nurse and doctor leading the research. They were happily interpreting the results, when he realized, to his horror, that the foil was upside-down; that is, that it needed turning over, and therefore the relationship between the target (output) and feature (input) was reversed. He sheepishly set it right, and in only seconds the head Doctor exclaimed, "That makes sense too!", and continued interpreting its (new and completely opposite) nuances.[1]

Humans are, and likely will remain, the best pattern-recognizers in existence – for the low dimensions in which we operate. But, we are perhaps too good; we tend to see patterns even when they don't exist. One of ERI's VPs, Dustin Hux, worked at the Virginia State Climatology Office while he was a grad student. A citizen sent in a videotape of purported cloud phenomena: "Could you weather experts explain this astonishing phenomena?". The un-narrated three-hour tape contained nothing but typical, summer (cumulus humulus) clouds. The citizen had seen "dragons in the clouds", where there (almost certainly weren't any.

A valuable step early in analysis is to seek to validate one's data internally: do the variables agree with one another? Finding as one team did on a corporate data set, that, "95% of the husbands are male", is useless in itself, but reveals something about the data's quality, provides audit questions and flags observations. Reliable analysis depends so strongly on the quality of the data that internal inconsistencies can hobble one's work; or, they can be clues to problems with the flow of information within the company and reveal a key process obstacle. ERI worked closely with a direct mail client years ago, and dove deeply into the data, looking for relationships between what was known about a potential customer and their resulting orders. We actually endangered our appearance of competence to the client by persisting to question the unexpectedly low numbers of catalogs being sent to some customers. Eventually, it was found that the "Merge/Purge house" was treating overseas purchasers the opposite of how they were instructed, and erroneously deleting them from the mailing lists – when they were actually some of the best prospects. This discovery was probably more helpful to the client's bottom line than most of our high-tech modeling work.[2]

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] Those who don't regularly research with computers seem to give more credence to their output, I've noticed. Perhaps like sausage being enjoyed most by those least familiar with how it's made.

[2] The analysis work though, combined with operational changes such as higher-quality catalog paper, did result in a doubling of the client's average sales per catalog within a year.

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