White Paper

It is a Mistake to…Listen Only to the Data

by John Elder

Founder & CEO

June 2014

ELDER RESEARCH

DATA SCIENCE & PREDICTIVE ANALYTICS

This article is Part 5 (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*.

Inducing models from data has the virtue of looking at the data afresh, not constrained by old hypotheses. But, while "letting the data speak", don't tune out received wisdom. Experience has taught this once brash analyst that those familiar with the domain are usually more vital to the solution of the problem than the technology we bring to bear.

Often, nothing *inside* the data will protect one from significant, but wrong, conclusions. Table 1 contains two variables about high school, averaged by state: cost and average *SAT* score (from about 1994). Our task, say, is to model their relationship to advise the legislature of the costs of improving our educational standing relative to nearby states. Figure 1 illustrates how the relationship between the two is significant – the Linear Regression *t*-statistic is over 4, for example, suggesting that such a strong relationship occurs randomly only 1/10,000 times.[1] However, the sign of the relationship is the opposite of what was expected. That is, to improve our standing (lower our *SAT* ranking), the graph suggests we need to reduce school funding!

Table 1: Spending and Rank of Average SAT Score by State

USA State	SAT Rank	$ Spent
AK	31	7877
AL	14	3648
AR	17	3334
AZ	25	4231
CA	34	4826
CO	23	4809
CT	35	7914
DC	49	8210
DE	37	6016
FL	40	5154
GA	50	4860
HI	44	5008
IA	1	4839
ID	22	3200
IL	10	5062
IN	47	5051
KS	6	5009
KY	18	4390
LA	16	4012
MA	33	6351
MD	32	6184
ME	41	5894
MI	20	5257

MN	3	5260
MO	13	4415
MS	12	3322
MT	19	5184
NB	8	4381
NC	48	4802
ND	2	3685
NH	28	5504
NJ	39	9159
NM	15	4446
NV	29	4564
NY	42	8500
OH	24	5639
OK	11	3742
OR	26	5291
PA	45	6534
RI	43	6989
SC	51	4327
SD	5	3730
TN	9	3707
TX	46	4238
UT	4	2993
VA	38	5360
VT	36	5740
WA	30	5045
WI	7	5946
WV	27	5046
WY	21	5255

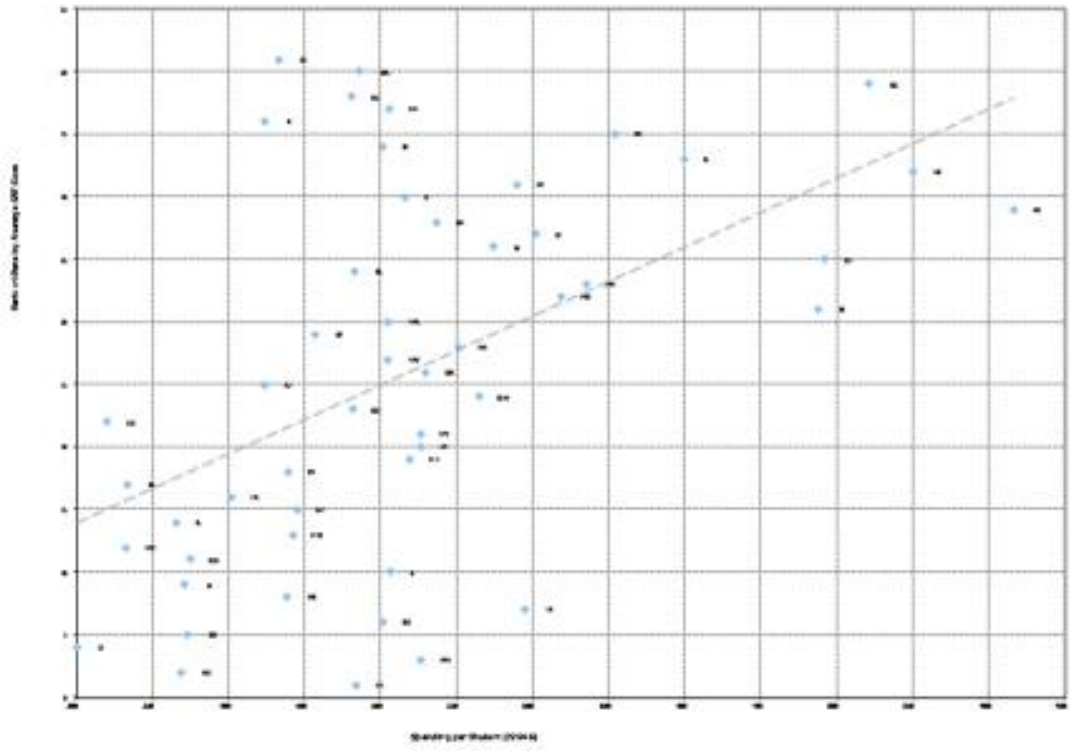

Figure 1: Rank of a State (in average SAT score) vs. its spending per student (circa 1994), and the least-squares regression estimate of their relationship

Observers of this example will often suggest adding further data – perhaps, for example, local living costs, or percent of the population in urban or rural settings — to help explain what is happening. But, the real problem is one of self-selection. The high-SAT/low-cost states are clustered mainly in the Midwest, where the test required for state universities (the best deal for one's dollar) is not the *SAT* but the *ACT*. Only those students aspiring to attend (presumably more prestigious) out-of-state schools go to the trouble of taking an extra standardized test, and their resulting average score is certainly higher than the larger population's would be. Additional variables in the database, in fact (other than proportion of students taking the SAT) would make the model more complex, and might obscure the fact that information external to the data is vital.

Observers of this example will often suggest adding further data – The above example employed typical "opportunistic", or found, data. But even data generated by a designed experiment needs external information. A national defense project from the early days of Neural Networks attempted to distinguish aerial images of forests with and without tanks in them. Perfect performance was achieved on the training set, and then also on an out-of-sample set of data that had been gathered at the same time but not used for training. This was celebrated but, wisely, a confirming study was performed. New images were collected on which the models performed extremely poorly. This drove investigation into the features driving the models and revealed them to be magnitude readings from specific locations of

the images; i.e., background pixels. It turns out that the day the tanks had been photographed was sunny, and that for non-tanks, cloudy![2] Even resampling the original data wouldn't have protected against this error, as the flaw was inherent in the generating experiment.

A second tanks and networks example, from my good friend and former colleague, Dean Abbott (who's got an excellent book out now!). Dean had worked at a San Diego defense contractor, where they sought to distinguish tanks and trucks from any aspect angle. Radars and mechanized vehicles are bulky and expensive to move around, so they fixed the radar installation and rotated a tank and a truck on separate large, rectangular platforms. Signals were beamed at different angles and the returns were extensively processed – using polynomial network models of subsets of principle components of Fourier transforms of the signals – and great accuracy in classification was achieved. However, seeking transparency (not easy for complex, multi-stage models) Dean discovered, much to his chagrin, that the source of the key distinguishing features determining vehicle type turned out to be the bushes beside one platform![3] Further, it is suspected that the angle estimation accuracy came from the signal reflecting from the platform corners – not a feature one will encounter in the field. Again, no modeling technology alone could correct for flaws in the data, and it took careful study of how the model worked to discover its weakness.

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] This theoretical result is confirmed by resampling using *Target Shuffling*; randomize the rankings and it takes about 10^4 tries before a correlation this strong is stumbled upon.

[2] PBS featured this project in a 1991 documentary series "The Machine that Changed the World": Episode IV, "The Thinking Machine".

[3] This excellent practice of trying to break one's own work, is so hard to do even if one is convinced of its need, that managers should pit teams with opposite reward metrics against one another in order to proof-test solutions.

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