White Paper

Evaluate the Validity of Your Discovery with Target Shuffling

John Elder
Founder

January 2014

ELDER RESEARCH
DATA SCIENCE · AI · MACHINE LEARNING

This paper is an excerpt from the article titled *"3 Ways to Test the Accuracy of Your Predictive Models"* published by Victoria Garment at PlottingSuccess.com.

John Elder is the founder of data mining and predictive analytics services firm Elder Research. He tests the statistical accuracy of his data mining results through a process called target shuffling. It's a method Elder says is particularly useful for identifying false positives, or when two events or variables occurring together are perceived to have a cause-and-effect relationship, as opposed to a coincidental one.

"The more variables you have, the easier it becomes to 'oversearch' and identify (false) patterns among them," Elder says—what he calls the 'vast search effect.'

As an example, he points to the Redskins Rule, where for over 70 years, if the Washington Redskins won their last home football game, the incumbent party would win the presidential election. "There's no real relationship between these two things," Elder says, "but for generations, they just happened to line up."

In situations like these, it becomes easy to make inferences that are not only incorrect, but can be dangerously misleading. To prevent this problem from occurring, Elder uses target shuffling with all of his clients. It's a process that reveals how likely it is that results occurred by chance in order to determine if a relationship between two variables is causal.

"Target shuffling is essentially a computer simulation that does what statistical tests were designed to when they were first invented," Elder explains. "But this method is much easier to understand, explain and use than those mathematical formulas."

Here's how the process works:

1. Randomly shuffle the output (target variable) on the training data to "break the relationship" between it and the input variables.
2. Search for combinations of variables having a high concentration of interesting outputs.
3. Save the "most interesting" result and repeat the process many times.
4. Look at a distribution of the collection of bogus "most interesting results" to see how much of apparent results can be extracted from random data.
5. Evaluate where on (or beyond) this distribution your actual results stand.
6. Use this as your "significance" measure.

Let’s break this down: imagine you have a math class full of students who are going to take a quiz. Before the quiz, everyone fills out a card with various personal information, such as name, age, how many siblings they have and what other math classes they’ve taken. Everyone then takes the quiz and receives their score.

To find out why certain students scored higher than others, you model the outputs (the score each student received) as a function of the inputs (students’ personal information) to identify patterns. Let’s say you find that older sisters had the highest quiz scores, which you think is a solid predictor of which types of future students will perform the best.

But depending on the size of the class and the number of questions you asked everyone, there’s always a chance that this relationship is not real, and therefore won’t hold true for the next class of students.

With target shuffling, you compare the same inputs and outputs against each other a second time to test the validity of the relationship. This time, however, you randomly shuffle the outputs so each student receives a different quiz score—Suzy gets Bob’s, Bob gets Emily’s, and so forth.

All of the inputs (personal information) remain the same for each person, but each now has a different output (test score) assigned to them. This effectively breaks the relationship between the inputs and the outputs without otherwise changing the data.

You then repeat this shuffling process over and over (perhaps 1,000 times), comparing the inputs with the randomly assigned outputs each time. While there should be no real relationship between each student’s personal information and these new, randomly assigned test scores, you’ll inevitably find some new false positives, or “bogus” relationships (e.g. older males receive the highest scores, women who also took Calculus receive the highest scores).

As you repeat the process, you record these “bogus” results using a histogram—a graphical representation of how data is distributed. You then evaluate where on this distribution (or beyond it) your model’s initial results (older sisters score highest) stand.

If you find that this result appears stronger than the best result of your shuffled data, you can be pleased with the original finding, as it was not matched by any random results.

If your initial result falls within the distribution, this tells you what proportion of random results did as well, which is your significance level. (For a study to be published in most medical journals, for example, the acceptable significance level is 0.05, meaning that there’s only a 5 percent chance of results that strong occurring by chance.)

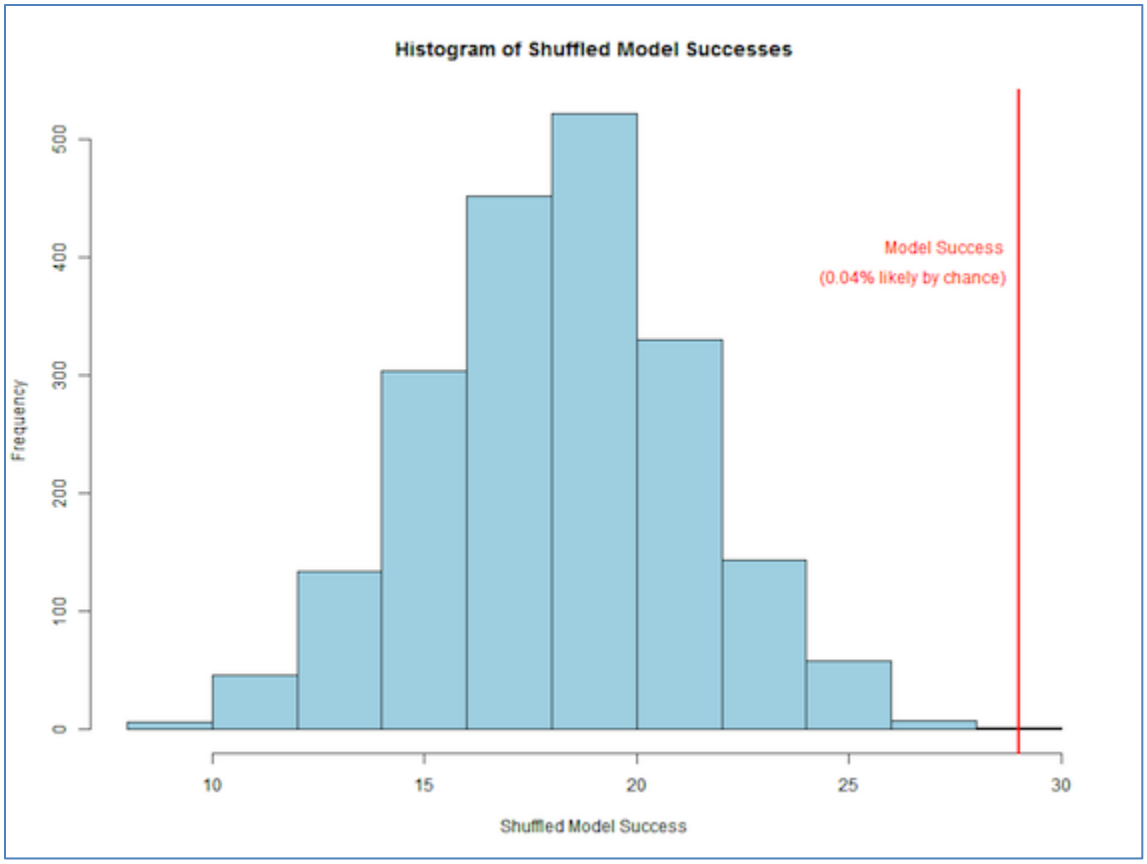

Figure 1: An example histogram comparing the success of a model to that of shuffled models

In the histogram pictured above, the model scored in the high 20's. Only 0.04 percent of the random, shuffled models performed better, meaning the model is significant to that level (and would meet the criteria of a publishable result in any journal).

John Elder first came up with target shuffling 20 years ago, when his firm was working with a client who wasn't sure if he wanted to invest more money into a new hedge fund. While the fund had done well in its first year, it had been a volatile ride, and the client was unsure if the success was real. A statistical test showed that the probability of the fund being that successful by chance was very low, but the client wasn't convinced.

So Elder performed 1,000 simulations where he shuffled the results (as described above) where the target variable was the buy or hold signal for each day. He then compared the random results to how the hedge fund had actually performed.

Out of 1,000 simulations, the random distribution returned better results in just 15 instances—in other words, there was a 1.5 percent chance that the hedge fund's success was a result of luck. This new way of presenting the data made sense to the client, and as a result he invested 10 times as much in the fund.

“I learned two lessons from that experience,” Elder says. “One is that target shuffling is a very good way to test non-traditional statistical problems. But more importantly, it’s a process that makes sense to a decision maker. Statistics is not persuasive to most people—it’s just too complex.

“If you’re a business person, you want to make decisions based upon things that are real and will hold up. So when you simulate a scenario like this, it quantifies how likely it is that the results you observed could have arisen by chance in a way that people can understand.”

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 · AI · MACHINE LEARNING