Three Pitfalls to Avoid in Predictive Modeling

By Marc Vincelli

Few would disagree with the power and promise of predictive modeling. From the Oakland A's use of predictive modeling to build a championship baseball team on a shoestring budget in 2002, to Google's use of search and text analytics to predict the H1N1 flu outbreak in 2009, the well-known examples of predictive modeling "successes" are numerous. Perhaps less widely recognized is the myriad of ways in which a predictive model can fail to perform as expected, often due to misconceptions or misrepresentations on the part of the analyst. In this article, I focus on three such pitfalls.

1. FORCING A PREDICTIVE MODEL ON A PROBLEM IN THE WORLD OF UNCERTAINTY

Economists and decision theorists have for some time distinguished between decisions made under risk and decisions made under uncertainty. In the world of risk, all alternatives, consequences and probabilities are known, or can be reasonably developed (using past experience, for example). In the world of uncertainty, some of this information is unknown, and possibly even unknowable.[1] While decision problems in the world of risk lend themselves well to statistical thinking, those in the world of uncertainty require good rules of thumb (heuristics)[2] and expert intuition balanced by deliberative reasoning.

The nuanced distinction between risk and uncertainty is important to consider when determining whether the predictive modeling toolkit is even appropriate for a given prediction problem. Some problems, such as predicting long-term interest rates or forecasting an individual's future financial needs, may involve too much uncertainty to appropriately leverage predictive modeling. In these cases, the application of professional judgment informed by a simulated fan of outcomes, in line with the RAND Corporation's robust decision-making (RDM) framework,[3] may be more prudent. Forcing a predictive model on a problem that resides within the world of uncertainty can result in suboptimal business decisions, a false sense of comfort and serious financial consequences. So before going too far down the predictive modeling path, the analyst is well advised to ask himself: "Am I dealing with a prediction problem in the world of risk, or uncertainty?"

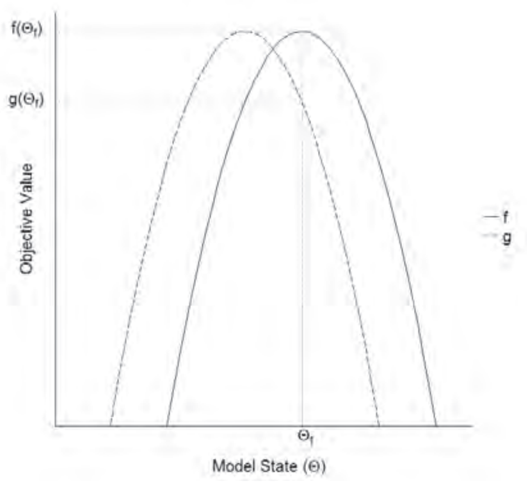
2. SUGGESTING THAT FUTURE MODEL RESULTS ARE LIKELY TO BE AS GOOD AS PAST RESULTS

Any model that has been optimized based on past data is likely to experience performance degradation upon implementation. This phenomenon, in which optimization produces a model that is more likely to perform well in the future but less likely to perform as well as past results suggest, has been called the *optimization paradox*.[4]

To see why future results are not likely to be as good as past results, consider the optimization of the objective function f in Figure 1. An objective function can be thought of as relating a quantity of interest (e.g., profit) to various model states (Θ) based on data available at the time the function was generated. Assume function f is the objective function we obtained just prior to model implementation. Under an optimization approach, we would implement the model state that maximizes (or minimizes, as appropriate) our objective function. Let us denote this optimized model state as Θ_f.

Now assume we have implemented the model and have accumulated more experience. Unless the future is just like the past, we can expect the emerging data to shift f in some unpredictable way, resulting in a new objective function g. How might the objective function shift? Well, in the neighborhood of the optimum $(\Theta_f , f(\Theta_f))$, which is our area of interest, the primary shifts would be up, down, left or right. Figure 1 illustrates the left-shifted case. The key to understanding the optimization paradox is to recognize that in three out of the four primary translations (i.e., shift down, shift left and shift right in our example), $g(\Theta_f)$ will be less than $f(\Theta_f)$. In other words, most of the time we should not expect future model results based on Θ_f to be as good as past results.

Figure 1: Illustration of Optimization Paradox with Left-Shifted Objective Function



Does this mean the analyst should avoid optimization? Absolutely not; optimization produces the "best" answer for a given objective function generated at a point in time, and a solution that will more likely than not continue to outperform its suboptimal alternatives. What it does mean, however, is that the analyst must appropriately set model performance expectations with the end-user. One way in which this can be done is to favor out-of-sample test results over in-sample test results when discussing expected performance.

3. OVER-SEARCHING TO FIND PATTERNS AND RELATIONSHIPS

One of the dangers with building predictive models on big data is over-searching, which can lead to spurious correlations and nonsensical models. If we dredge through enough data, we will eventually—by chance alone—find something that appears to be correlated to our target variable but really has no relationship whatsoever. It is incumbent on the analyst to apply his own professional judgment to validate the inclusion of variables and to avoid testing a hypothesis on variable inclusion with evidence used in constructing the hypothesis itself.

Perhaps one of the best known examples involving spurious correlation is the Super Bowl Indicator, which "predicted" that when a premerger National Football League team won the Super Bowl, the U.S. stock market would rise, and when an old AFL team won the Super Bowl, the U.S. stock market would fall. It turns out that between 1967 and 2013, this indicator was correct more than 70 percent of the time. Surprisingly, the indicator was even discussed in the highly respected *Financial Analysts Journal*.[5] So would you be willing to put your money and/or reputation on the line that this correlation is predictive? Only one's good judgment, and not a model, can answer that question.

CONCLUSION

As powerful and promising as predictive modeling can be, practitioners have a responsibility to ensure that the toolkit is applied appropriately and that end-users understand each model's "sphere of competence" (including intended usage, expected performance and risks). Three steps one can take toward this end are to:

- avoid applying predictive modeling to problems that reside within the world of uncertainty,
- explain to the end-user that future model results are unlikely to be as good as results optimized to the training data, and
- identify and exclude variables with spurious correlations. ■

Marc Vincelli, ASA, M.Sc., is a principal consultant with Fortis Analytics in Kitchener-Waterloo, Canada. He can be reached at *marc_vincelli@fortisanalytics.com.*

ENDNOTES

[1] Pablo A. Guerron-Quintana, "Risk and Uncertainty," Philadelphia Fed Business Review Q1 (2012): *https://www.phil.frb.org/research-and-data/publications/business-review/2012/q1/brq112_risk-and-uncertainty.pdf.*

[2] Gerd Gigerenzer, *Risk Savvy: How to Make Good Decisions* (New York: Penguin Books, 2014).

[3] Robert J. Lempert, et al., "Making Good Decisions Without Predictions: Robust Decision Making for Planning Under Deep Uncertainty," *RAND Corporation Research Briefs* RB-9701 (2013): *http://www.rand.org/pubs/research_briefs/RB9701.html.*

[4] Curtis M. Faith, *The Way of the Turtle* (New York: McGraw-Hill, 2007).

[5] Robert R. Johnson, "Is It Time to Sack the 'Super Bowl Indicator'?" Total Return (blog), *The Wall Street Journal*. Jan. 22, 2014, *http://blogs.wsj.com/totalreturn/2014/01/22/is-it-time-to-sack-the-super-bowl-indicator.*