Prediction is a Young Science

Lisa R. Goldberg*

forthcoming in Notices of the American Mathematical Society

January 9, 2016†

Predictions go back at least as far as the Oracle of Delphi and the astrologers of the Chinese Han Dynasty. Arguably, they are as old as the human race. However, our recently acquired capability to collect, store and analyze data has emphasized statistics and elevated prediction to a science that pervades virtually everything we do. Despite recent advances, prediction is a young science. We are just beginning to explore its limits.

1 Scientific prediction can be humbling and confusing

Time to eat some crow
-Nate Silver

On July 8, 2014, Germany won the World Cup with a stunning 7–1 victory over Brazil. This was a problem for superstar statistician Nate Silver, whose Poisson distribution-based Soccer Power Index (SPI) model had forecast a win for Brazil, and had set the odds of Germany beating Brazil by 6 goals or more at 1 in 4,000.

Silver has been using statistical models to predict outcomes of sporting events since the early 2000s, but he is best known for predicting elections. In 2008, Silver and his colleagues at FiveThirtyEight correctly called every state in the US presidential election with the exception of Indiana and the 2nd congressional district in Nebraska, which awards its own electoral vote. In 2012, FiveThirtyEight correctly called the US presidential election in all fifty states (including the toss-up in Florida) and the District of Columbia. Silver’s book on statistics, The Signal and The Noise, became an international bestseller.1 “Triumph of the Nerds” and other articles celebrating data science appeared, exceptionally, in the mainstream media.

In response to questions about SPI’s disastrous Brazil-Germany prediction, Silver explained,

---

*Consortium for Data Analytics in Risk, Departments of Economics of Statistics, University of California, Berkeley, CA 94720-3880, USA, email: lrg@berkeley.edu.

†This article has benefitted from conversations with Bob Anderson, Jeff Bohn, Alan Cummings, Nick Gunther, Ola Mahmoud, Barry Mazur, Caroline Ribet, Ken Ribet, Stephanie Ribet, Alex Shkolnik, Philip Stark, and Roger Stein.

Statistical models can fail at the extreme tails of a distribution. There often isn’t enough data to distinguish a 1-in-400 from a 1-in-4,000 from a 1-in-40,000 probability.

Statistical models rely on estimated distributions, outcomes and likelihoods, to make predictions. The tails of a distribution are populated by rare or never-before-seen outcomes. It is virtually impossible to assess the credibility of the estimated likelihood of a rare event.

Silver was not the only one to mis-forecast the outcome of the 2014 World Cup. Many of the big investment banks had used statistical methods to predict a win for Brazil. The Financial Times demoted data science from hero to goat with headlines like “Brazil’s World Cup eludes banks’ best minds.” On the other hand, Microsoft was delighted to point out that its prediction engine, Cortana, had correctly predicted the outcome in 15 of 16 World Cup matches including the Brazil-Germany final.

2 Extreme events may seem obvious in hindsight

On June 22, 2007, the investment bank Bear Stearns was quietly working to rescue two of its hedge funds: the Bear Stearns High-Grade Structured Credit Fund and the Bear Stearns High-Grade Structured Credit Enhanced Leveraged Fund. These funds were trading securities backed by sub-prime mortgages, which involve loans to homeowners by banks who readily acknowledge the possibility that some of the homeowners might default. However, relatively few considered the possibility that lots of borrowers might default at the same time.

The rescue failed since the sub-prime market had begun to collapse in response to a rapid rise in mortgage payment delinquency. By mid-July, Bear Stearns disclosed that the two funds had lost nearly all their value. Smooth and upward-trending equity markets began to rock. This was the preamble to the Global Financial Crisis that dominated the ensuing two years, and which has been blamed for the destruction of trillions of dollars, the loss of millions of jobs, and the suicides of more than 10,000 individuals.

Despite plentiful data, the elaborate and expensive risk management systems in place at many financial institutions in 2007 failed to alert investors to the impending crisis. There were exceptions. In The Big Short, Michael Lewis tells the tales of a few investors who noticed that housing prices were inflated, acted on their observations, and profited. For the vast majority of us, however, the crisis signals were obscured by the noise of everyday life.

2 According to Robinson (2014), “For banks, football predictions are a bit of fun, as well as a way to demonstrate the statistical savvy and forecasting prowess of research departments typically focused on interest rates and equities.”

3 In March 2009, Bear Stearns was acquired by JP Morgan at a fire sale price.

4 More information about suicides related to the Global Financial Crisis is in Aaron et al. (2014).

5 The caliber of risk management differed across financial institutions. According to https://en.wikipedia.org/wiki/Subprime_crisis_impact_timeline#2007, Goldman Sachs claimed after the fact that it had noticed the risks in the sub-prime mortgage market and it began to reduce its exposure in December 2006.
I recall a San Francisco limousine driver telling me early in 2007 about the houses he was flipping in Georgia. It seemed odd, but I did not ask myself what it implied. I could have known. That nagging I-could-have-known feeling that comes in the aftermath of a disaster is an example of hindsight bias, and it depends on our selective memory of what turned out to matter. In his bestselling book, *Thinking, Fast and Slow*, Daniel Kahneman explains the behavioral roots of hindsight bias. These include the availability heuristic, our tendency to rely in information that easily comes to mind, and the representativeness heuristic, our tendency to profile.

3 A correct prediction can also be a bad prediction

The Monty Hall Problem goes like this:

There is a sports car behind one of three closed doors, but the other two doors hide goats. You get to select a door and keep whatever is behind it. After you make an initial choice, Monty Hall opens one of the two unselected doors, revealing that it does not hide the sports car. He offers you the opportunity to switch to the other unopened door or stay with your original choice. Should you stick or switch?

If you stick with the door you originally selected, you may be lucky and win the car. But the probability that the car is behind the door you originally selected is 1/3, while the probability that car is behind the unselected door Monty Hall failed to open is 2/3. The prediction that the car is behind the door you originally selected is bad, even if it turns out to be correct.

A skilled predictor may never make a bad prediction, but she will inevitably make some predictions that turn out to be incorrect. This conundrum is concisely summarized in the Nassim Nicholas Taleb’s Table of Confusion, which pairs terms that tend to be mixed up: luck and skill, belief and knowledge, theory and reality, signal and noise, randomness and determinism.

4 Scientific prediction learns from mistakes

In 1948, major polls in the US famously predicted that New York governor Thomas Dewey would beat Harry Truman in the presidential election.

---

6My review of Kahneman (2011) is in Goldberg (2013).
7The foundation of behavioral heuristics is Kahneman and Tversky (1979).
8The Monty Hall Problem is explained at https://www.youtube.com/watch?v=4Lb-6rZxx0 and https://www.youtube.com/watch?v=ugbWqWcxrg&feature=youtu.be&t=2m32s.
9The Table of Confusion on Taleb (2005, Page 3) lists “luck and skills,” rather than “luck and skill.”
The predictions were based on quota sampling, which targets subsets of a population that are pre-specified rather than representative. The predictions were also stale. As George Gallup, co-chairman of the Gallup organization explained, “We stopped polling a few weeks too soon.” Modern election polls use random sampling rather than quota sampling, and they run up to the last minute.

Today, scientific predictions tell us which movies, books and restaurants we will enjoy, what diseases will afflict us in the future, whether a cancerous growth on one of our feet will prove fatal, how drugs will interact,\(^\text{10}\) who will be our soul mates, and when we might be wiped out by rising sea levels and ocean acidification. It can be difficult to distinguish good predictions from bad ones, since a correct prediction need not be good. A prediction might be good if it is based on ample data and an algorithm that learns,\(^\text{11}\) if it fits in-sample and is validated out-of-sample, and if it is interpretable and appealing to common sense. It may be impossible to test predictions of the time and location of the next big earthquake, or the efficacy and safety of a cure for a rare disease, or the time and cause of the next financial crisis.

But some scientific predictions have demonstrated track records. Advanced warnings of hurricanes and tornadoes have allowed communities at risk to evacuate before storms hit. Airline travel is remarkably safe, in part, due to accurate weather forecasts. When Google Navigation tells you that, despite some traffic, you will reach your destination by 3:12 pm, there are reasons to believe it.

References


\(^{10}\)More information about data science and drug interactions is in Tatonetti et al. (2012).

\(^{11}\)Bayesian models update prior distributions with new information as it becomes available.


