Aaron Brown

Superforecasting

Reflecting on Tetlock and Gardner's "Ten Commandments for Aspiring Superforecasters"

Philip Tetlock's *Expert Political Judgment* (Princeton University Press, 2005) is one of the most important books of recent decades. People who haven't read it remember one of its points – that forecasts by experts are no better than random guesses, and in some definable circumstances, worse. The more famous the expert, the less accurate the results. This was based on an exhaustive 20-year study, and the statistical evidence was overwhelming.

People who have read the book know that it contained much deeper and more interesting insights. For one thing, the book explores in careful philosophic terms exactly what it means for a forecast to be accurate. The many reasons people give that inaccurate forecasts can be valuable are refuted with both logic and empirical investigation. Another interesting topic is a subtle investigation of why and how expert predictions are pushed away from truth. But the most important revelations in the book are that some people are good forecasters, beating not only random chance, but also simple extrapolation algorithms and prediction market prices, and that forecasting skill can be taught.

That last result captured the attention of the little-known Intelligence Advanced Research Project Activity (IARPA); its cousin, the Defense Advanced Research Project Agency (DARPA), is more famous, particularly for DARPANet, the precursor to the Internet. IARPA funded a prediction tournament pitting amateurs recruited by Tetlock and his grad students against four other academic teams – which used methods such as prediction trading markets and machine learning algorithms – as well as government analysts (who, among other advantages, have access to classified information). The questions were selected by IARPA as ones of greatest interest to policy makers, such as: "*Who will be in control of Yemen at the end of 2015?*" with forecasters asked to assign probabilities to answers such as: "*President Hadi will be restored to power*," or "*The civil war will be continuing*." Of course, the questions and answers are qualified carefully to minimize ambiguity.

Tetlock's teams did so much better than the other teams that, in a loss to science, the experiment was called off, and Tetlock devoted his energies to improving his methods. It would have been useful to learn if other groups could build on Tetlock's success and add additional accuracy from their own methods. The scale of the victory is impressive, 30 percent better, as measured by the Brier score. In a binary question, that translates to assigning 20 percent less probability, on average, to the wrong answer. If Tetlock's teams assigned 60 percent probability to the right answer and 40 percent to the wrong, the competition averaged 52 percent to the right answer and 48 percent to the wrong. That could be because Tetlock's teams were right more often, or because they were more decisive in their predictions (you could get the 60 percent average figure by assigning 100 percent probability to an answer and being right 60 percent of the time, or always having the right answer but only assigning 60 percent probability, or anything in between) or some combination.

Now, Tetlock has a new book out, with Dan Gardner as a coauthor (I reviewed Gardner's *The Science of Fear* in the May 2009 issue of this magazine). *Superforecasting* describes the experiment and introduces us to some of the amateurs who beat the experts and the other academics. It's a fascinating account, with a lot of useful information, but I'm going to skip to the end, to discuss the authors' "ten commandments for aspiring superforecasters". Here's their version of everything you need to outpredict the best and the brightest of US intelligence agencies, as well as my thoughts on each commandment.

Triage

This is straightforward enough. Don't waste analytical firepower on things predictable enough with simple rules, and don't spin your wheels attacking problems when you'll never get useful traction.

While this is sensible, it can be hard to apply; sometimes the only way to figure out how predictable something is, is to try to predict it.

But there are times when this rule is very useful. When I first got interested in basketball betting in the 1970s, there were quants who tried to analyze the basketball game itself from scratch. That seemed hard to me, compared to asking which team was likely to attract more betting interest. As Los Angeles is a rich and high-betting city, and as the Lakers were a glamorous team, it wasn't hard to guess that the betting public would disproportionately favor the Lakers and therefore the spread would be slanted against the Lakers. "*Bet against the Lakers at home*," took a lot less mental effort than simulating basketball games.

Break seemingly intractable problems into tractable subproblems

Enrico Fermi was famous for recommending this approach. To use a different example from the book, suppose you want to know the chance that the Democrats will take control of the US Senate in the 2016 elections. Unless you follow elections closely, you probably have little intuition about that question. I did a Google search (2016 US Senate elections Democrat odds) and, after eliminating duplicated information and unclear answers, six sites suggested that the chance of a Democratic victory was negligible, three said it was unlikely, and one said it was possible but less than 50 percent. So, if you didn't weight the credibility of the sites or assess their arguments, you'd probably think the chances were pretty low.

But now let's break the problem down. We can look up the composition of the Senate: it's 54 Republicans, 44 Democrats and two Independents. There are ten Democrats and 24 Republicans up for re-election in 2016. Historically, a decent simple model is that a party has two chances in three of having a 90 percent chance of retaining its seats, and one chance in three of having a 60 percent chance, and that the party's two draws are independent, and each race is independent, given the conditional probabilities.

In that case, the Democrats' main hope of picking up five or more seats is to get a good draw (90 percent chance in its contested seats), the Republicans to get a bad draw (60 percent chance in their contested seats), giving Democrats a 95 percent chance of victory. But this scenario only has a 22 percent likelihood. The next best shot comes if both parties get bad draws, a 65 percent chance that Democrats win, and an 11 percent likelihood of this scenario. If both parties get good draws, the Democrats' chances are only 4 percent; this scenario has a 44 percent likelihood. I think people who underestimate the Democrats' chances are overweighting this scenario. After all, it's the most likely, and it gives the Democrats little chance. Finally, in the 22 percent of the time that the Democrats get a bad draw and the Republicans get a good draw, the Democrats win only 0.3 percent of the time. Combining all these numbers gives an overall 30 percent chance that the Democrats will pick up five or more seats and take control of the Senate.

Obviously, I wouldn't bet on my number, certainly not against someone who had more political knowledge and had done more work. But I trust it over the nonquantitative, careless reasoning of most commentators. I broke the problem down into parts: how many seats have to reverse parties for the Democrats to take control; what is the chance the Democrats have a good election; what is the chance the Republicans have a good election; and, given those draws, what is the chance of a Democratic victory in the Senate? I don't have great confidence in any of my answers, but I don't think they're so far wrong that the chances of a Democratic victory could be less than, say, 5 percent, as the predictions I found with Google suggested.

Strike the right balance between inside and outside views

Inside views are specific to the problem; outside views take a step back. For example, in the Senate question above, I took an inside view, trying to analyze the question based on directly relevant data. For an outside view, I might ask: "*When most popular commentators think an election result is unlikely, how often are they surprised?*" Now, instead of looking at US Senate election data, I'd look at a broader range of elections and pay attention to the prior predictions. Obviously, this would be an improvement to my analysis.

This is another way of stating a popular technique in risk management. If you want to estimate the center of a probability distribution, say, from the 5th to the 95th percentile, you rely on specific, recent, quantitative data. If you want to estimate the tails, say, beyond the 1st or 99th percentile, you look instead at general, long-term, qualitative data. Thus, the volatility measured over the last two weeks using intraday data gives a pretty good estimate of tomorrow's volatility of a liquid security most of the time; but if I want to know what might happen if tomorrow's move is more than three standard deviations, I'm better off looking at big moves in general over history, rather than what this security did recently.

Strike the right balance between under- and overreacting to evidence

One of the findings in the book is that the best forecasters were frequent incremental updaters. As new evidence came in, they tweaked their estimates. On occasion, however, they would make large, sudden changes in forecast. Most people ignored new information, but would occasionally make large forecast changes in response to mild new evidence.

I think the key here is that good forecasters have a story or model. As information comes in, the story stays the same, but the constituent probabilities adjust somewhat. For example, a poll showing that Democrats were losing favor with likely Senate voters might cause me to adjust my probability of the Democrats getting a bad draw from 33 percent to 40 percent. That would cause a small adjustment to my forecast, from a 30 percent to 29 percent chance of a Democratic victory. Someone without a story wouldn't know how to calibrate the information, so they'd likely either ignore it or make a larger change in forecast.

But the real value of a story is that when something happens that should be impossible in your model, you throw everything out and start over. Even if you get to a similar quantitative forecast, the new model means you'll have a different reaction to future data. In my case, suppose some information came out that linked the probability of a good Democratic draw to a bad Republican draw. Although most people think about elections this way, I don't see it in the historical data. It seems that nationwide Democratic areas can tire of Democrats more or less independently of whether nationwide Republican areas are tiring of Republicans. Disaffection with one party does not seem to imply affection for the other. If I'm wrong about that, my entire approach goes out the window, and I'd need to build a new model taking account of that.

Look for the clashing causal factors at work in each problem

One of the most striking findings in the book is that you can judge forecast quality by word choices. Good forecasters say: "*on the other hand*," "*nevertheless*," and "*but*" a lot. Bad forecasters prefer: "*moreover*," "*in addition*," and "*even more*." Good forecasters bob and weave to a conclusion; bad ones believe in: "*Damn the torpedoes, full speed ahead*."

The challenge here is to make sure you have accounted for the opposing factors without letting your forecast soften into irrelevance. Another kind of bad forecaster has so many qualifications and evasions that there is no forecast at all.

Surprisingly, to many people, the superforecasters were not only more accurate than experts, but also more decisive, giving forecasts farther away from 50 percent, on average. So, mind the clashing factors, but don't be afraid to give the edge to the strong one.

Strive to distinguish as many degrees of doubt as the problem permits, but no more

This is a classic dilemma, as illustrated in *Star Trek*:

Kirk: "*Do you think Harry Mudd is down there, Spock?*"

Spock: "*The probability of his presence on Motherlode is 81 percent plus or minus 0.53.*"

McCoy: "*Why can't you just say Mudd's probably there?*"

Spock: "*I just did, doctor.*"

Kirk asks for a forecast, and Spock answers with a precisely distinguished one. Note that he is not only giving precise numbers, but he is also even distinguishing between the objective probability and his uncertain estimate of it, rather than convoluting the two into a single number. McCoy objects that Spock has distinguished more degrees of doubt than the problem permits. Spock demurs.

I take the middle here, traditionally where Kirk stands. I think Spock is overprecise, and McCoy underprecise. I'd go with, "About 4 to 1 in favor, Captain."

Strike the right balance between under- and overconfidence, between prudence and decisiveness

The hard part about this one is that confidence is negatively correlated to accuracy. Even experienced risk takers bet more when they're wrong than when they're right; and the most confident people are generally the least reliable.

Another problem is that the degree of confidence and decisiveness is often predetermined by the situation rather than something a forecaster can choose. Sometimes you get put on the spot and are forced to choose between the Lady or the Tiger. Sometimes you are only allowed to express slight differences of opinion with the official verdict.

The solution to the first problem is to keep careful, objective records, preferably by a third party. The solution to the second is to improve communication and decision making so that nuanced opinions and divergent opinions can be expressed and integrated, and so that the loudest voice doesn't automatically carry the day.

The authors are being a bit sneaky with this one. One of the ways that Tetlock got such good performance was to first combine forecasts of different people, then push the results away from 50 percent. So, he built a system that led to accurate but underconfident results, and he added confidence artificially.

Look for the errors behind your mistakes but beware of rearview-mirror hindsight biases

It's not just mistakes – even successful forecasts could be improved (and even incorrect ones usually have some threads of truth). Rigorous review is essential. If you only look at results, it takes too long to make improvements. You have to break things down to your subproblems and see which ones were reasonable approximations and which ones were not, or if there was an outside factor you didn't consider that messed things up.

One form of hindsight bias is to react to an unexpected occurrence by predicting that it will happen again. Nassim Taleb wrote that people underestimate the probability of novel extreme events, but overestimate the probabilities of recent extreme events. Another form is to react to error by losing all confidence in prediction, and just move all forecasts toward 50 percent (even worse is to react to success by overconfidence).

Unless you were completely wrong, and the actual outcome was one with almost no probability in your model, there should be pieces of your reasoning to preserve. And unless you were completely right, there should be pieces to improve.

Bring out the best in others and let others bring out the best in you

There is definitely a wisdom of crowds, if you have the right crowd (diverse backgrounds, skills, and opinions) and aggregation structure (brings out each person's independent knowledge and integrates them into a consensus forecast). In my experience, the most important factor is incentives. If you can structure things so that everyone wins from the group's successful forecast, and everyone loses from the group's failure, it's relatively easy to get people to work together constructively. But if people have personal agendas, it's difficult to do better than the best individual forecaster, and you may well end up worse than the worst individual forecaster.

Master the error-balancing bicycle

I think what the authors really mean here is get out and forecast. A lot of the commandments involve balancing competing principles. That's easy to say, but hard to do when you first try. However, like riding a bicycle, there comes a magic point where it switches from impossible to easy and natural. Or else you end up bruised, battered, and near your starting point.

Assuming you ever do get the knack, and I think most people can in surprisingly little time, the remaining dangers are all overcompensation. You need to balance a lot of things to ride a bicycle. This will be a test for the superforecasters in the book. What happens now that they're famous? Can Tetlock's teams retain the humility and independence that beat out all the other forecasting methods, or will they succumb to forecaster hubris? Will they be content to ride their bicycles down quiet country lanes, or will they come to grief trying to outrun cars on urban highways?