risks

MDPI

Book Review

Superforecasting: The Art and Science of Prediction. By Philip Tetlock and Dan Gardner

Daniel Buncic

University of St. Gallen, Institute of Mathematics and Statistics, Bodanstrasse 6, 9000 St. Gallen, Switzerland; daniel.buncic@gmail.com; Tel.: +41-71-224-2604

Academic Editor: Ola Mahmoud
Received: 15 June 2016; Accepted: 16 June 2016; Published: 5 July 2016

Let me say from the outset that this is an excellent book to read. It is not only informative, as it should be for a book on forecasting, but it is highly entertaining. It provides a review of some of the worst forecasts in history, and also gives a fairly personal account of a number of "*ordinary*" individuals that Philip Tetlock got to know rather well during various forecasting tournaments and research projects that he was involved in [1]. I highly recommend the book not only to practitioners who are dealing with forecasting problems on a regular basis, or academics that are interested in more technical, research questions related scoring rules, the evaluation of forecasts, and the like, but also to a much wider audience. In fact, the book's opening sentence "*We are all forecasters*" sums it up nicely that each and everyone of us, either formally or informally, consciously or subconsciously, is engaged in (some sort of) forecasting on a daily basis. The topic of the book is thus timely and relevant for everyone. I particularly liked the adopted writing style of the two authors, Philip Tetlock and Dan Gardner. The book is structured coherently. Moreover, it is written like a novel (rather than a textbook) in the sense that it gets the reader hooked by telling the story of the book in an unfolding fashion, with glimpses of the results of Tetlock's most recent forecasting project, the Good Judgment Project (GJP), revealed slowly and progressively in stages throughout the book. Needless to say, the combination of a Professor of psychology and a journalist as authors of the book, indeed make it difficult to put the book down once you get started.

So what is the book about? Well, it is about forecasting, but there are many such books [2]. What makes the book different from most other standard treatments on forecasting is that it gives a detailed account of the forecasting performance of a large number of "*ordinary*" individuals that volunteered to take part in various forecasting competitions that Tetlock has organized since about 1984. These "*ordinary*" individuals, forecasting laymen so to speak, would compete with professional, government as well as academic, forecasters on a variety of geopolitical and also economic topics of interest [3]. Moreover, the book describes the reasoning and the thought process that these so called "*Superforecasters*" employ to arrive at a probabilistic prediction of the event of interest.

What are "*Superforecasters*"? Tetlock and Gardner describe them as ordinary, everyday individuals, that are logical thinkers with an eagerness to "*share their underused talents*" [1]. Superforecasters have a keen interest in understanding the mechanism behind things, as opposed to just getting accurate forecasts without knowing why. They question their line of reasoning. They are willing to test new ideas. They never settle on one and only one forecasting mechanism, but rather remain in a "*perpetual*

[1] On the first page of the book, Bill Flack, a retired government employee of the US Department of Agriculture in Arizona is introduced, who took part in Tetlock's forecasting tournaments and who has achieved the status of a "*Superforecaster*".
[2] The most recent book in an economic or econometric setting is by [2].
[3] The types of questions forecasters would have to answer were questions like: "Will India or Brazil become a permanent member of the UN Security Council in the next two years?" or "Will North Korea detonate a nuclear device before the end of this year? and the like. Participants (forecasters) would have to give a probabilistic forecast, that is, the likelihood of the event occurring.

beta" mode, realizing that life, as well as the world around them, is complex, that nothing is for certain, that everything is constantly evolving. Moreover, they are well aware of the fallacies and tricks that the mind can play on us (humans) to "*make something look certain*", leading to the temptation to come to the "*obvious*" conclusion too quickly, without second guessing it, and re-analyzing the evidence that is presented over and over again.

A good part of the book deals with outlining the typical character traits that Superforecasters have. In this context, Tetlock and Gardner try to emphasize and convince the reader that "*forecasting is not a 'you have it or you don't' talent*", but that Superforecasting can be learned. Part of the book's purpose is thus to illustrate to the reader how to become a Superforecaster. On page 191, under the heading of "*Pulling it all Together*", Tetlock and Gardner list some specific key characteristics that a typical (modal) Superforecaster possesses, using the following four headline points to group them:

(1) philosophic outlook,
(2) abilities and thinking style,
(3) method of forecasting, and
(4) work ethic.

There are a number of themes in the book that I particularly liked and found extremely relevant, but let me highlight only a few here. I have an econometric or model based forecasting background. Although the type of forecasting that is described in this book is quite different, as it is based on subjective or judgmental forecasting, I see many parallels between the key characteristics that Superforecasters possess when constructing judgmental forecasts and those utilized by Econometricians, when building statistical prediction models. For instance, Tetlock and Gardner talk about "*Foxes*" being better forecasters than "*Hedgehogs*", with "*foxes knowing many little things, while the hedgehog knows one big thing*." Since conditions in the real world are rather complex and constantly changing, Tetlock and Gardner find in their research that it is better to move away from an *"I know one big thing"* approach to forecasting, and rather adopt an *"I know many little things"* approach. In an econometric (model based) forecasting world, this means that one needs to move away from a single model (or predictor) framework, to one that utilizes instead a potentially large number of (small) models, which are then combined or averaged to produce a single forecast of the target variable of interest. Such a modelling strategy is fairly common today (see for instance [3], for an application of this approach in the equity premium forecasting literature, or [4,5] in the context of exchange rate forecasting and equilibrium credit modelling).

Although the gains from a *"diversified"* prediction framework, where the average of the predictions from many individual forecasts (i.e., "*the wisdom of the crowds*") is used, seem obvious, Tetlock and Gardner discuss at length the main obstacles they faced when forming "*Superteams*", that is, teams of Superforecaster to implement such a strategy in the GJP forecast competition (see Chapter 9 and also the discussion that follows). There were two main concerns. The first was "*groupthink*", that is, everyone in the group converging to the *"average views"* of the group, without questioning the logic or motivation behind the arguments to come to that conclusion, or the sources of information used to form the prediction. The second was rancor and dysfunction. The overall conclusion from this chapter is that consensus is not always good, and disagreement not always bad. Moreover, one should focus on constructive confrontation. What is interesting to point out here is that a similar loss in the forecast gains from model averaging can be encountered in an econometric forecast combination framework, if the individual forecasts that make up the aggregate are too similar.

An illustrative visual example that highlights this issue is seen from Figure 1 below, which shows forecasts of the US unemployment rate 12 months into the future from the Survey of Professional Forecasters (thin light blue lines), together with the actual realized unemployment rate (thick dark red line). There are a number of features that are intriguing from this plot, but let me discuss just a few. First, note that the overall predictions from the professional forecasters have the same trajectory, that is, turning points in the series, despite showing some variation in them. The predictions are dispersed, so there is variability in the forecasts of the individuals, but the turning points no one seems to be

able to predict. This is particularly evident around 2001 and 2008. Second, the forecasts lag by about 1 year (4 quarters) behind. This suggests that all professional forecasters seem to utilize, to some extent at least, the same type of dynamic econometric model to generate the forecasts. Third, all prediction models that are used by professional forecasters seem to be based on the idea of mean reversion, where the series today is constructed as a weighted average of the (unconditional) mean of the series and the deviation of the series from its mean. If the mean of the series was higher in the recent past, the prediction from the model will tend to revert back to this higher mean. This property of the model based forecasts is clearly and systematically visible from Figure 1. For instance, at the beginning of the sample, the mean was between 6 and 7. Over the period from 1995 to 2001, nearly all forecasts made are systematically above the realized value. Similarly, between 1999 to 2001, the mean had dropped, resulting in systematic under-predictions of the unemployment rate from 2001 to 2004. This pattern continues. Combining forecasts from such similar individual predictions is not going to lead to a great deal of improvement, and can hence be viewed as the econometric or model based analogue to "*groupthink*".

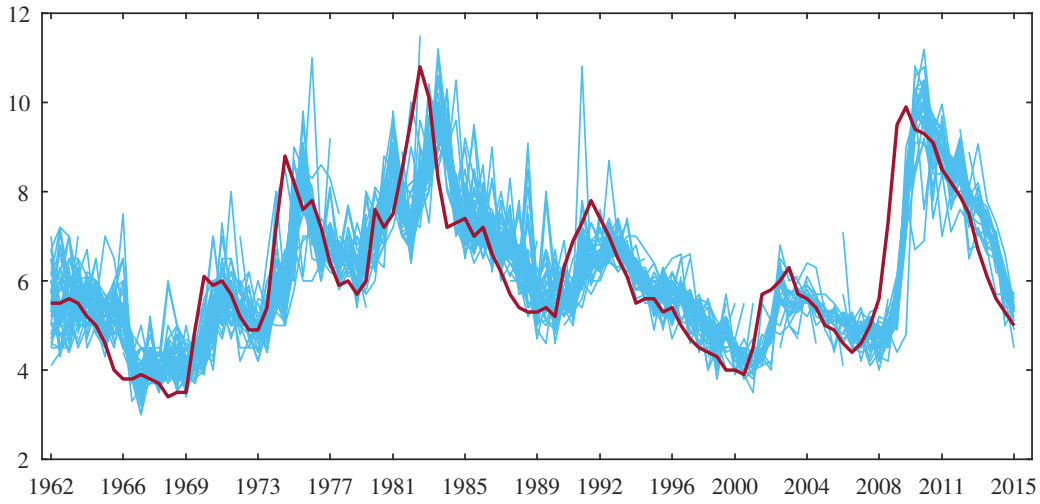

Figure 1. Forecasts of the US unemployment rate 12 months ahead from the Survey of Professional Forecasters (SPF). The thin light blue lines are the predictions from the forecasters, and the thick dark red line is the realized unemployment rate. Data were obtained from: https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/.

Since this is a critical review, I should say a few words about what I did not like so much in the book. All assessments of forecast accuracy and comparisons across forecasters in the book (and the GJP as a whole) are based on the Brier Score (BS) [4]. Evidently, when wanting to assess how well forecasters are doing, one will need to settle on a performance measure. The Brier Score is a proper score function designed to measure the accuracy of probabilistic forecasts. It is thus perfectly suitable for the type of predictions that are considered in the book and the GJP. The Brier Score is formally defined as:

$$\mathrm{BS} = 2\Big[\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2\Big], \tag{1}$$

where N denotes the number of predictions that are made, $\hat{p}_i$ is the predicted probability of event i, and y_i is the realisation or outcome of the event, where y_i takes the value of 1 if the event occurred, and

[4] BS seems to be the commonly used abbreviation for the Brier Score (see for instance https://en.wikipedia.org/wiki/Brier_score). I am thus using it here in accordance with standard practice in the literature, rather than as an expression of contempt or dislike with the measure.

0 if it did not [5]. A BS of 0 means that the forecaster is doing really well, in fact, as well as is possible, correctly predicting the binary outcome every time, while a BS of 2 means the forecaster is getting it wrong every time. A BS of 0.5 is the score you would get from random guessing (the dart-throwing chimp metaphor in Tetlock and Gardner), or alternatively, assigning a 50% probability to every event that is forecasted.

So what is the problem here? Because the Brier Score uses a quadratic loss function, the penalty from getting forecasts very wrong increases non-linearly. With getting forecasts very wrong I mean the scenario of assigning a predicted probability $\hat{p}_i$ of 1 (or close to one) for an event to occur, but the realised value y_i is 0 (the event did not occur). Or conversely, my prediction being the event is impossible to occur (i.e., $\hat{p}_i = 0$ or close to 0), yet it occurs. These are grave forecasting errors and there is nothing wrong with penalizing them harshly. It makes sense to do that from a utility based motivation as well. However, if I am aware of this as a forecaster (and I am sure many forecasters are), I will be more inclined to provide weak predictive signals, in the sense that I will have a preference for my forecasts to be close to the 50% uninformative prediction most of the time, unless I feel extremely confident. This is exactly the type of forecast that nobody wants to have. Think of the predictions that the President's advisors were providing regarding the compound in Pakistan, discussed at length in Chapter 6 of the book under the heading "The third setting". On page 135 it says: *"... the median estimate of the CIA officers — the "wisdom of the crowd" — was around* 70%. *And yet Obama declares the reality to be "fifty-fifty"*. Thus, the use of scoring rules that encourage weak predictions provide weak information for policy relevant actions. In an econometric or model based prediction setting, obtaining forecasted probabilities that are nearly always close to 50% would raise doubts about the informativeness of the model, or, to put it in econometrics jargon, the goodness of fit of the model to the data. Moreover, the predictor variables that go into the model, i.e., the conditioning variables used to generate the forecast, would be considered as irrelevant for the outcome variable.

Another issue with how BS is used throughout the book is that only point estimates are provided, without any mention of the uncertainty surrounding these estimates. Because the outcomes y_i are realisations of a random variable, the Brier Score in (1), which depends on y_i, will also be a random variable. Making statements of the form: "*their collective Brier score was 0.25, compared with 0.37 for all other forecasters*" on page 94 is uninformative, as there is no measure of variation reported in the difference between the two groups' Brier scores [6]. From a statistical point of view, one is inclined to ask: "Is this difference in BS significant?" or "What are the confidence intervals of the two scores?" In theory, these can be computed quite easily; the authors of [6], in fact, derive standard errors and confidence intervals for the Brier Score and Brier Skill Score. However, one crucial assumption in the derivation (see the move from line 2 to 3 in equation 16 on page 994 in [6]) is that the pair of predicted probabilities and outcomes of events are random, i.e., independent across i (the events to be predicted). In general, this may not seem like a strong assumption to make. But in the given context, it would depend on the type of forecast questions that are being asked and would require us to think about questions like "What is the likelihood of two (or more) of the predicted events occurring jointly?" If the set of questions to be answered in the forecast competition include: "Will the Kurdistan Regional Government hold a referendum on national independence this year?" and "Will either the Swiss or French inquiries find elevated levels of polonium in the remains of Yasser Arafat's body?", then it probably is easy to argue that these two events can be regarded as independent of each other. However, consider the case where the set of questions includes: "Is Britain going to leave the European Union?"

---

[5] Note that y_i is thus a realisation from a Bernoulli random variable Y_i, with success probability p_i. Also, Tetlock and Gardner use the original scaled (multiplied by 2) version of BS (they define the lower and upper bounds of BS to be 0 and 2 on page 93).

[6] I should say here in defence of the authors that, first, they are probably very well aware of this, given their background, and second, that they do report results of the Brier Score from various years of the same Superforecaster being ranked, which gives more credence to the difference not just being due to chance. Nevertheless, the issue still remains.

and "Is the Pound Sterling going to depreciate by 20% against the Euro over the next 3 months?", or "Are borrowing costs in Britain going to increase over the next 6 months", and the like. Then, it becomes much harder to justify the assumption of independence of the events being predicted. Computing the variance of BS is thus more complicated in practice.

As Nelly Furtado said (sang actually) and many people before her, "All good things come to an end". So let me conclude this review with the words that the book is a great read, very informative, entertaining, and, above all, provides the reader with a summary of the experiences of forecasters in a life long project that Tetlock has been involved in since about 1984.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Tetlock, P.; Gardner, D. *Superforecasting: The Art and Science of Prediction*; Crown: New York, NY, USA, 2015.
2. Elliott, G.; Timmermann, A. *Economic Forecasting*; Princeton University Press: Princeton, NJ, USA, 2016.
3. Rapach, D.E.; Strauss, J.K.; Zhou, G. Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy. *Rev. Financ. Stud.* **2010**, *23*, 821–862.
4. Buncic, D.; Piras, G.D. Heterogenous Agents, the Financial Crisis and Exchange Rate Predictability. *J. Int. Money Financ.* **2016**, *60*, 313–359.
5. Buncic, D.; Melecky, M. Equilibrium credit: The reference point for macroprudential supervisors. *J. Bank. Financ.* **2014**, *41*, 135–154.
6. Bradley, A.A.; Schwartz, S.S.; Hashino, T. Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score. *Weather Forecast.* **2008**, *23*, 992–1006.