Prediction is not everything, but everything is prediction

Leonardo Egidi

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'Bruno de Finetti', Università degli Studi di Trieste, Trieste, Italy.

E-mail: legidi@units.it

Jonah Sol Gabry

Department of Statistics, Columbia University, New York, USA.

E-mail: jgabry@gmail.com

Abstract. Prediction is an unavoidable task for data scientists, and over the last decades statistics and machine learning became the most popular 'prediction weapons' in many fields. However, prediction should always be associated with a measure of uncertainty, because from it only we can reconstruct and falsify the model/algorithm decisions. Machine learning methods offer many point-predictions, but they rarely yield some measure of uncertainty, whereas statistical models usually do a bad job in communicating predictive results. According to the Popper's falsificationism theory, natural and physical sciences can be falsified on the ground of wrong predictions: though, for social sciences this is not always true. We move then to a weak instrumentalist philosophy: predictive accuracy is not always constitutive of scientific success, especially in social sciences.

Keywords: Prediction; Popper's falsificationism philosophy; Weak instrumentalism; Predictive accuracy; Machine learning

1. Introduction

As motivated by the falsificationism approach (Popper, 1934) and many philosophers of science, prediction has a primary role in the progress of science; however, this is often a controversial argument—see Kuhn (1962) and Lakatos (1976) for some criticisms. Popper argues that theories, in order to be scientific, must be falsifiable on the ground of their predictions: wrong predictions should perhaps push the scientists to reject their theories or to re-formulate them, conversely exact predictions should corroborate a scientific theory. Popper's philosophy is instrumentalist in a strong sense (Hitchcock and Sober, 2004) when applied to physical and natural sciences: predictive accuracy is constitutive of scientific success, not only symptomatic of it, and prediction works as a confirmation theory tool for science.

Since the 1940s, with the growing availability of fast computers and the use of simulation routines, science expanded its boundaries and extended the existing frameworks in new directions; think, for instance, at the Manhattan project in Los Alamos, when the problem of neutron diffusion in fissionable material allowed Stanislaw Ulam and Nicholas Metropolis to invent and develop Markov Chain Monte Carlo Methods through the ENIAC computer. In particular, the birth and the growth of probabilistic and statistical methods have made the 'debut of science in society' possible, whereas the growing ability of data and the development of sophisticated

social sciences. In Section 3, we focus on prediction for statistical learning, whereas the weak instrumentalist philosophy is detailed in Section 4. Section 5 proposes an applied example for the football Russia World Cup 2018, whereas Section 6 concludes.

2. Prediction for science or science for prediction?

2.1. It is prediction part of the science design?

The main stages required to formulate a scientific law are summarized by Russell (1931) as follows: (1) observation of some relevant facts; (2) formulation of a hypothesis underlying and explaining the facts above; (3) deduction of some consequences from this hypothesis. In his opinion, the modern scientific method is born with Galileo Galilei, father of the law of falling bodies, and with Johannes Kepler, who discovered the three laws of planetary motion:

> Scientific method, as we understand it, comes into the world full-fledged with Galileo (1564-1642), and, to a somewhat lesser degree, in his contemporary, Kepler (1571-1630). [...] They proceeded from observation of particular facts to the establishment of exact quantitative laws, by means of which future particular facts could be predicted.

Then, the law of universal gravitation of Isaac Newton embodied the two previous theories, whereas the theory of the general relativity of Albert Einstein generalized the Newton's theory. Thus, in the last 500 years, physics—and, more generally, science—advanced by falsification and generalization of the previous theories, by providing new and more exciting theories to predict new natural facts and highlighting the confirmation nature of prediction. In general, as Hitchcock and Sober (2004) argue, mathematical descriptions of the invariant behaviour of a physical phenomenon are essentially predictive: further experiments and observations can validate these theories.

However, the link of prediction with the scientific laws is in our opinion more ambiguous than what people are usually inclined to think. The following questions arise: is prediction a central step in science? Is prediction a relevant aim of science? A negative answer to the first question could be seen in disagreement with some *instrumentalist* scientists, who would claim that, from an instrumental perspective, predictive success is not merely *symptomatic* of scientific success, but it is also *constitutive* of scientific success (Hitchcock and Sober, 2004). A more sophisticated answer could be: prediction is not explicitly part of the formulation of a scientific hypothesis (1)–(3) *at the time the law is posed*, but it becomes relevant and relevant as science advances; the chain of events which brought Newton to generalize the theories of Galilei and Kepler first, and Einstein to revisit the gravitational law of Newton then, was supposedly based on the fallacy of some predictions, and it gained sense only *ex-post*. The fact that the bodies in proximity to the earth surface were revealed by Newton to not fall exactly with a constant acceleration—the acceleration slightly rises as they get closer to the earth—did not make the Galilei's law of constant acceleration for falling bodies less scientific, or totally wrong from a scientific point of view. Scientific falsification detected by wrong predictions (Popper, 1934) is a powerful and exceptional tool, but along this paper we feel to warn about its abuse/misuse.

Over the last decades, scientific predictions became popular not only in the context of physics and natural science, but for social sciences as well. Steps (1)–(3) above are widely used by social scientists and statisticians to build consistent theories about human and social behaviours:

227), but gaining lower voters' percentage (46.1% vs 48.2%). According to various online poll aggregators, Hillary Clinton was given a 65% or 80% or 90% chance of winning the electoral college. As Gelman (2016b) argues:

> These probabilities were high because Clinton had been leading in the polls for months; the probabilities were not 100% because it was recognized that the final polls might be off by quite a bit from the actual election outcome. Small differences in how the polls were averaged corresponded to large apparent differences in win probabilities; hence we argued that the forecasts that were appearing, were not so different as they seemed based on those reported odds. The final summary is that the polls were off by about 2% (or maybe 3%, depending on which poll averaging youre using), which, again, is a real error of moderate size that happened to be highly consequential given the distribution of the votes in the states this year.

In November 2016, many modelers, included Nate Silver, the founder of the well-known FiveThirtyEight blog (`https://fivethirtyeight.com`), failed to predict the Trumps' win. However, it is naive to conclude that those models failed because their underlying mechanism was wrong; rather, political science predictions cannot entirely act as theory's confirmation tools, due to many reasons attributed, for instance, to nonresponse and voters' turnout, as explained by Gelman (2016a):

> Yes, the probability statements are not invalidated by the occurrence of a low-probability event. But we can learn from these low-probability outcomes. In the polling example, yes an error of 2% is within what one might expect from nonsampling error in national poll aggregates, but the point is that nonsampling error has a reason: its not just random. In this case it seems to have arisen from a combination of differential nonresponse, unexpected changes in turnout, and some sloppy modeling choices. It makes sense to try to understand this, not to just say that random things happen and leave it at that.

3. The role of prediction in statistical learning

3.1. From the observed to the observable

As statisticians, we are often faced with a double task: first, creating a sound mathematical model to accomodate the data and retrieve useful inferences for our parameters—at the time being, we make no distinction here between classical and Bayesian inference; second, using this model to make predictions, and this is rarely accounted by the statisticians in a transparent way.

However, statistics has always been thought as the *science of inference*, or *science of estimates*, and inference is always seen as separate from prediction. Inference is based on an underlying mathematical model for the data-generating process (Bzdok et al., 2018), its main task is to describe an unknown mechanism working through generalization: the inferential laws should in fact be as broad as possible, ideally valid for the population of interest, and not symptomatic of the observed data (it is out of the scope of this paper to review the distinct inferential approaches). Prediction moves from the observed to the unobserved, being the action designed to forecast future events without requiring a full understanding of the data- generation process.

bias can be sometimes bogus, and does not entirely reflect the needs of the statistician. Rather, if we are framed in a Bayesian context we intend the unobserved values $\tilde{y}$ to come from the posterior predictive distribution, denoted here by $p(\tilde{y}|y)$, which incorporates the intrinsic uncertainty propagating from the parameters—summarized by the posterior distribution—to the observable future values. Through this quantity, we could define an expected predictive density (EPD) measure for a new dataset. In much previous literature about predictive accuracy, such as the Akaike Information Criterion (AIC) (Akaike, 1973), there is not any link to the model's uncertainty, since the measure of model's accuracy is evaluated conditionally on parameters' points estimates. The Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) is a sort of AIC Bayesian version, in which the maximum likelihood estimate is replaced by the posterior mean for the parameters, and the number of parameters is replaced by a measure of effective number of parameters.

Recent proposals such as the Watanabe-Akaike Information Criteria (WAIC) (Watanabe, 2010) and Leave-One-Out cross validation Information Criteria (LOOIC) (Vehtari et al., 2017) go in the direction of data granularity, by definition of the expected log pointwise predictive density for a new dataset (ELPPD). These approaches require the computation of the log-pointwise predictive density $p(\tilde{y}_i|y)$ for each new observable value $\tilde{y}_i$. Of course, the true distribution is unknown, and this measure has to be approximated, for instance via leave-one-out cross validation.

Although all the predictive information criteria may fail in some practical situations, LOOIC and WAIC offer the possibility to provide a measure of predictive accuracy based on the single data points, in a computationally efficient way (both the methods are implemented in the `loo` R package (Vehtari et al., 2019)). Despite not conclusive for the predictive accuracy of a statistical model, these techniques allow in many situations to compare distinct models by the acknowledgement of an intrinsic uncertainty propagating from the parameters to the observable future values: in such a viewpoint, *observable values, and not parameters, are really relevant.* A transparent predictive tool should encompass data, parameters and future data all together: in such a way, the falsification of a single piece makes the joint model falsifiable. In Section 4, we make this point even more clear.

3.3. The two cultures

As brilliantly argued by Breiman et al. (2001), there are two cultures in the use of statistical modeling to reach conclusions from data: a stochastic data model consisting of predictors, parameters and random noise to explain the response variable y is adopted by the data modeling culture; a function of the predictors to predict the response variable y is assumed by the algorithmic modeling culture, also named machine learning (ML) culture. The two approaches strongly differ in their validation: goodness-of-fit tests vs. predictive accuracy on out-of-sample data. It is evident that the data modeling culture—linear regression, generalized linear models, Cox model, etc.—is aimed at extracting some information about how nature is associating the response variable to the dependent variable, whereas the algorithmic culture—decision and classification trees, neural nets—is more oriented to predict future values of the response variable given the values of the predictors.

In the mid-1980s neural nets and decision trees became incredibly popular (Breiman et al., 1984) in areas where parametric data models were not applicable, such as speech recognition, image recognition, handwriting recognition, and prediction in financial markets. In analysing

framework. The number of predictors at each split of a random forest is a tuning parameters fixed at $\sqrt{p}$ in most cases, but in practice the best values for these parameters will depend on the problem. Predictions should corroborate or reject an underlying theory, but if the method (the theory) is tuned and selected on the ground of its predictive accuracy, the theory to be falsified is bogus, and not posed in a transparent way.

As statisticians and (data) scientists, demanded to build models for social and physical sciences, our efforts should be addressed to produce good, transparent and well posed algorithms/models, and make them falsifiable upon a strong check (Gelman and Shalizi, 2013). Our skepticism regards the role of prediction in falsifying our models, for such a reason we would claim to be *weak instrumentalists*: predictions and predictive accuracy are a central task of science, but only sometimes they are constitutive of scientific success.

In other way said, a supposedly valid scientific theory should exist *before* the future data have been revealed, and produce some immediate benefits to the scientific community, similarly as the falling bodies theory of Galilei first, and the law of universal gravitation of Newton then: corroborating or rejecting a model/algorithm on the basis of observable future values only is often far from the scientists' requirements and economic funds of the current project.

4.2. *The falsificationist Bayesianism framework: going beyond inference and prediction*

Gelman and Shalizi (2013) argue that a key part of Bayesian data analysis regards the model checking through posterior predictive checks. In such a view, the prior is seen as a testable part of the Bayesian model and is open to falsification: from such intuition, Gelman and Hennig (2017) name this framework *falsificationist Bayesianism*.

As stated by Gelman et al. (2013), the process of Bayesian data analysis can be idealized by dividing it into the following three steps:

(a) Setting up a full probability model–a joint probability distribution—for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.

(b) Conditioning on observed data: calculating and interpreting the appropriate posterior distribution, i.e. the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.

(c) Evaluating the fit of the model and the implications of the resulting posterior distribution: how well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step (a)? In response, one can alter or expand the model and repeat the three steps.

In the above paradigm, predictions are never mentioned. But this does not mean that predictions are not relevant in the Bayesian paradigm. Denoted by $\tilde{y}$ the unobserved vector of future values, we may derive the posterior predictive distribution as

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta, \tag{2}$$

where $p(\theta|y)$ is the posterior distribution for θ, whereas $p(\tilde{y}|\theta)$ is the likelihood function for future observable values. Equation (2) may be resambled in the following way:

* * *

* * * Do you agree with this? Anyway, we could partially rephrase. I would like to connect this concept with the concept you stated: there is ~~now~~ something, in statistics, that is not limited and done with the idea of prediction?

Let (y_n^H, y_n^A) denote the observed number of goals scored by the home and the away team in the n-th game, respectively. A general bivariate Poisson model allowing for goals' correlation (Karlis and Ntzoufras, 2003) is the following:

$$\begin{aligned} Y_n^H, Y_n^A | \lambda_{1n}, \lambda_{2n}, \lambda_{3n} &\sim \text{BivPoisson}(\lambda_{1n}, \lambda_{2n}, \lambda_{3n}) \\ \log(\lambda_{1n}) &= \theta + \text{att}_{h_n} + \text{def}_{a_n} + \frac{\gamma}{2} w_n \\ \log(\lambda_{2n}) &= \theta + \text{att}_{a_n} + \text{def}_{h_n} - \frac{\gamma}{2} w_n \\ \log(\lambda_{3n}) &= \beta_0, \end{aligned} \tag{4}$$

where the case $\lambda_{3n} = 0$ reduces to the double Poisson model (Baio and Blangiardo, 2010). $\lambda_{1n}, \lambda_{2n}$ represent the scoring rates for the home and the away team, respectively, where: θ is the common baseline parameter; the parameters att_T and def_T represent the attack and the defence abilities, respectively, for each team T, $T = 1, \ldots, N_T$; the nested indexes $h_n, a_n = 1, \ldots, N_T$ denote the home and the away team playing in the n-th game, respectively; the only predictor is $w_n = (\text{rank}_{h_n} - \text{rank}_{a_n})$, the difference of the FIFA World Rankings (`https://www.fifa.com/fifa-world-ranking/`)—expressed in FIFA ranking points divided by 10^3—between the home and the away team in the n-th game, multiplied by a parameter $\gamma/2$. This last term tries to correct for the well-known phenomenon of *draw inflation* (Karlis and Ntzoufras, 2003), favouring the draw occurrence when teams are close in terms of their FIFA rankings. The value of the FIFA ranking difference w included in the models was considered on June 7th, only a bunch of days before the tournament takes place. In a Bayesian framework, attack and defence parameters are usually assigned some noninformative prior distributions (Baio and Blangiardo, 2010) and imposed a sum-to-zero constraint to achieve identifiability.

We decided to train our statistical models/ML techniques on distinct portions of matches from the group stage, where teams are more heterogeneous in terms of their FIFA rankings and actual strengths. To assess predictive performance between statistical models and ML algorithms in predicting football outcomes, we compare the double Poisson and the bivariate Poisson model, fitted by `rstan` package (Stan Development Team, 2018), with five ML procedures: Random Forest, Classification and Regression Trees (CART), Bagged CART, Multivariate Adaptive Regression Splines (MARS) and Neural Network, according to their standard use as provided by the `caret` package (Kuhn, 2019). The three different prediction scenarios are:

A *Train* 75% of randomly selected group stage matches
Test Remaining 25% group stage matches

B *Train* Group stage matches
Test Knockout stage

C *Train* Group stage matches for which both the teams have a Fifa ranking greater than 1
Test Knockout stage.

Figure 1 displays for each scenario the values for the FIFA rankings for the training set matches (blue points) and the test set matches (orange points), along with the line Rank 1 = Rank 2,

implying that the ranking difference is $w = 0$. In Scenario A, the test set matches are randomly selected from the group stage, and they do not show any particular pattern around the line $w = 0$. In scenarios B and C, test set matches belong to the knockout stage, where the teams are expected to be stronger and closer each other in terms of their rankings. In fact, the majority of the orange points (13 out of 16) is displayed towards the bottom right corner—higher rankings—and closer to the line $w = 0$—closer strengths. Scenario B uses more and more data to predict test set results—all the 48 group stage matches—whereas Scenario C only six matches.

Figure 2 depicts the posterior predictive distribution (p5 and p7) of the number of goals scored by France and Croatia during the final from the bivariate Poisson model. Darker regions are associated with higher probabilities, whereas the red squadre is in correspondence with the observed result, 4-2. From this plot, one could be tempted to conclude that the bivariate Poisson model completely failed to predict the match; however, the global probability of France win within the 90 minutes—obtained summing the single probabilities over the lower triangle of the plot—is about 42%, against the 29% chance of win for Croatia (p1 and p2). Fro this plot only we can acknowledge the intrinsic variability in our model predictions (p5).

To have a glimpse about statistical and ML procedures' predictive performance, Table 2 shows the accuracy in the predictions for the seven methods and the three scenarios. Assuming that higher predictive accuracies should not entirely suggest the best scientific methods (p1), we analyse the performance of the methods by focusing on pro and cons. As suggested by Figure 1a, Scenario A is the most noisy in terms of rankings' differences, being its test set constituted by matches randomly chosen from the group stage, without any kind of pattern. As it is intuitive, ML techniques (Random Forest and Neural Nets), perform better, since they 'shake' the training set (p8) in such a way to retrieve the highest predictive accuracy. The ML performances dramatically decrease in Scenario B and C, where learning from the training set should be focused on predicting the knockout stage. ML algorithms learn less and in a very random way, but it is not clear why (p10). As already argued, the choice of the training and the test set can dramatically change the predictive performance of the ML algorithms, which over-perform statistical models only when considering a portion of the group stage to predict the remaining group stage matches. Should maybe we conclude that statistical models are better scientific tools to predict the World Cup? Not at all (p1), but we can learn from this example to improve over the next World Cups (p4).

By concluding, from this simple case-study we cannot openly falsify our statistical/ML techniques on the ground of future predictions. However, Poisson models seem to be less sensitive to the training set structure, and then falsifiable in a broader sense.

6. Discussion

Prediction is central in the progress of science and became even more relevant in statistics and data science, as the availability of new computational tools became common to accomodate data and predict new events. The entire field of science changed a lot over the last decades, new disciplines entered in the scientific gotha, and social sciences became new frontiers where predictive accuracy was strongly required.

Natural and physical sciences progressed by means of Popper's falsificationism philosophy, whose one of the main consequences is the strong predictivism: scientific theories should be fal-

Gelman, A. (2016a). Election surprise, and three ways of thinking about probability.

Gelman, A. (2016b). Explanations for that shocking 2% shift.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A. and C. Hennig (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 180*(4), 967–1033.

Gelman, A. and C. R. Shalizi (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology 66*(1), 8–38.

Hitchcock, C. and E. Sober (2004). Prediction versus accommodation and the risk of overfitting. *The British journal for the philosophy of science 55*(1), 1–34.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Volume 1, pp. 278–282. IEEE.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*, Volume 112. Springer.

Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician) 52*(3), 381–393.

Kuhn, M. (2019). *caret: Classification and Regression Training*. R package version 6.0-84.

Kuhn, T. S. (1962). The structure of scientific revolutions. *Chicago and London*.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can theories be refuted?*, pp. 205–259. Springer.

Maher, P. (1988). Prediction, accommodation, and the logic of discovery. In *PSA: Proceedings of the Biennial meeting of the philosophy of science association*, Volume 1988, pp. 273–285. Philosophy of Science Association.

Popper, K. (1934). *The logic of scientific discovery*. Routledge.

Popper, K. (1944). The poverty of historicism, ii. a criticism of historicist methods. *Economica 11*(43), 119–137.

Popper, K. (1945). The poverty of historicism, iii. *Economica 12*(46), 69–89.

Russell, B. (1931). *The scientific outlook*. Routledge.

Sarewitz, D. and R. Pielke Jr (1999). Prediction in science and policy. *Technology in Society 21*(2), 121–133.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology) 64*(4), 583–639.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2.