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.

1. Introduction

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. Russell argues that to perfectly apply the scientific method, we should collect some facts A, B, C, D and, by induction, formulate a general law, referred to in A, B, C, D are examples. If this general law is true, we should then retrieve the same facts by deduction. 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.

As Russel states, Galilei provided a generalized version of the theory by considering only a few observations—further experiments confirmed the appropriateness of his hypothesis—whereas Kepler had formulated his theory by looking at the planets' motions. Isaac Newton (1642-1726) brought the theories of Galilei and Kepler together in his encompassing law of universal gravitation. When Albert Einstein (1879-1955) generalized the Newton's law in his theory of general relativity, the world was shocked by a sort of *final theory* about the universe. 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.

However, the link of prediction with the scientific laws is in our opinion more ambiguous than what people are usually inclined to think. Is prediction a central step in science? A naive

answer could be: no, prediction is not explicitly part of the formulation of a scientific hypothesis (1)–(3), as drawn by Russell. Is prediction a relevant aim of science? A naive answer could be: yes, scientists formulate quantitative laws, 'by means of which future particular facts could be predicted', and can then validate the goodness of their assumptions by somehow measuring the predictive accuracy.

The first answer 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 scientifically wrong. Scientific falsification detected by wrong predictions (Popper, 1934) is a powerful and exceptional tool, but 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. Many algorithms and models have been developed to predict political scenarios, policies effects, sport results, fluctuations of national Gross Domestic Product (GDPs), and many others. The role played by prediction for social sciences is more obscure (Popper, 1944, 1945) and much controversial than for natural science, though data scientists are every day more and more asked to build 'weapons of mass prediction' in many social contexts. The way in which they formulate their theories underlying some data follows in the majority of the circumstances the scheme outlined by Bertrand Russell and reported at the beginning of this section: the stage of 'hypothesis formulation' is vague here, but may be interpreted either in form of the classical statistical testing procedure, or in terms of a model to be checked. Though, the actual outcome may be far away from the predictions: Trump's win in the 2016 US Presidential Elections, Brexit, and Leicester's Premier League's win were very low-probability events, but they occurred. Can all of these rare events falsify the finest algorithms and models designed to not predict their occurrence? Our naive and tentative answer is no, they can't. On the other hand, a statistical procedure that had foreseen Leicester winning the Premier League at the beginning of the season 2015-2016 would have been a very bad model.

2.2. Prediction as a confirmation theory approach

For Popper (Popper, 1934), a theory is scientific only if is falsifiable, where the falsification of a theory is meant to be the the possibility to compare its predictions with the observed data. In his view, theories whose predictions conflict with any observed evidence must be rejected: prediction corroborates (or confirms) a theory when it survives an attempt at falsification; prediction delegitimizes a theory when it does not pass the falsification test.

The confirmation nature of prediction is crucial in natural sciences, such as physics. In general, as Hitchcock and Sober (2004) argue, mathematical descriptions of the invariant behaviour of a physical phenomenon—such as Newton's and Keplero's laws, or Maxwell's equations—are

essentially predictive: further experiments and observations can validate these theories.

A well-known historical example of predictive confirmation in chemistry dates back to the middle of the 19th century—see Maher (1988) for a detailed version of the example. At that time, more than 60 chemical elements were known, and new ones continuing to be discovered. Some prominent chemists attempted to determine their atomic weights, density and other properties, by collecting many experimental observations. In 1871, the Russian chemist Dmitri Mendeleev noticed that arranging the elements by their atomic weights, valences and other chemical properties tended to show a periodical recurrence. He found some gaps in the pattern, and he argued that these missing values corresponded to some existing elements which had not yet been discovered: he named three of these elements eka-aluminium, eka-boron, and eka-silicon, by giving some detailed description of their properties. Despite the skepticism of the scientific community, the French Paul-Emile Lecoq de Boisbaudran in 1874, the Swedish Lars Fredrik Nilson in 1878, and the German Clemens Winkler in 1886 discovered three elements which corresponded to descriptions of eka-aluminium, eka-boron, and exa-silicon, respectively: these three elements are better known now as gallium, scandium and germanium. The predictive ability of Mendeleev was remarkable—the Royal Society awarded him the Davy Medal in 1882—, and the new discovered elements well represent pieces of evidence which confirmed the theory.

Predictive confirmation is still ambiguous in social sciences. As argued by Popper (1944, 1945) and Sarewitz and Pielke Jr (1999), social sciences have long tried to emulate physical sciences in developing invariant mathematical laws of human behaviour and interaction to predict economics quantities, elections, policies, etc.; many scholars agreed about the fact that a social theory should be judged on its power to predict (Friedman, 1953).

However, we believe that social science predictions require more and more motivations to validate the underlying theory. In the 2016 United States Presidential Election the Republican Donald Trump defeated the Democrat Hillary Clinton by winning the Electoral College (304 vs 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. Statistics

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. Each person is more or less confident with the weather's predictions or with presidential election predictions, but rarely that person is aware of the underlying statistical model required to produce that forecast, unless he is a statistician/data scientist. In such a view, inference seems hard and obscure, and prediction easy and close to the people. This is often a paradoxical argument, since the inference is often associated to the *explanation* of the problem, and should be relevant and available to the majority of the population.

As statisticians, we are often faced with a double task. First, we must create a sound mathematical model to accomodate the data and retrieve useful inferences for our parameters—there is not distinction here between classical and Bayesian statistics, they are both designed to draw inferential conclusions, either in form of point estimates/confidence intervals or in terms of posterior quantiles/credibility intervals. Then, we should use this model to make predictions, but this is rarely accounted by the statisticians in a transparent way. First of all, should the statistician use all the n data to build a reasonable/useful model, or could he take only a portion of the sample to accomodate the model (the training set) and use the remaining values to validate the model (the test set)? This apparently naive question pushed many scholars to debate about the supposed supremacy of prediction over accomodation (Maher, 1988; Hitchcock and Sober, 2004; Worrall, 2014). According to this philosophical point of view, the statistician should ask himself whether he wants models that are true—or approximately true—or predictively accurate.

It is well-known that the performance of a statistical method requires low variance as well low bias. Suppose we use a set of observations $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ to fit (or train) the statistical model $Y = f(X) + \varepsilon$, where the single x_i is the observed value of the predictor/covariate/independent variable X_i, the single y_i is the observed value of the dependent/response variable Y_i, f is an unknown mathematical function of X, and ε is the random error. Fitting a model means producing an estimate $\hat{f}$ for the unknown f: in the univariate linear regression case

$f(X) = \beta_0 + \beta_1 X$, this is translated in estimating the parameters β_0, β_1 by finding $\hat{\beta}_0, \hat{\beta}_1$. Let (x_{n+1}, y_{n+1}) be a new observation not used to train the model, we can then define the expected test mean square error as:

$$E\left[y_{n+1} - \hat{f}(x_{n+1})\right]^2 = \text{Var}(\hat{f}(x_{n+1})) + \text{Bias}^2(\hat{f}(x_{n+1})) + \text{Var}(\varepsilon), \tag{1}$$

where $\hat{f}(x_{n+1})$ is the prediction for y_{n+1} produced by $\hat{f}$, $\text{Var}(\hat{f}(x_{n+1}))$ is the variance for $\hat{f}(x_{n+1})$, $\text{Bias}^2(\hat{f}(x_{n+1}))$ is the squared bias, and $\text{Var}(\varepsilon)$ is the error variance. More complex models, with lower bias, tend to overfit the data, by yielding poor predictive results and then higher variance; conversely, too simple models tend to not fit the data adequately and have higher bias. Statistical procedures often incur in the bias-variance trade-off (James et al., 2013), the challenge is to find a compromise by controlling both the bias and the variance.

When building a model for real-life applications to extract information from the data, it is good practice to keep in mind the bias-variance trade-off. Nevertheless, it is often problematic to assess the performance of a statistical model by looking at the elements in Equation (1). When f is unobserved, it is even impossible to compute the expected test MSE.

In our practice, prediction should not be assimilated to 'take a rabbit out of a hat', but looking at its inherent uncertainty. Splitting the predictions' uncertainty in variance and squared bias can be sometimes bogus, and does not entirely reflect the needs of the statistician. Rather, we intend the unobserved values $\tilde{y}$ to come from a probability distribution, denoted here by $p(\tilde{y}|y)$, such that 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—summarized by their posterior distribution—to the observable future values—through the posterior predictive distribution. In Section 4, we make this point even more clear. A transparent predictive tool should encompass data, parameters and future data all together, in such a way that the falsification of a single piece makes the joint model falsifiable.

3.2. Machine learning

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 real data from these fields, the only criterion to evaluate these algorithms was predictive accuracy: this is translated in finding an algorithm $f(x)$ able to be a good predictor for y for future values of x, the so called *test set*.

Data scientists are used to train their procedures on the *training set*, which is chosen at the beginning in many possible ways. A common strategy is to select the first half of a dataset to train the algorithm, and the second half to test it. Another strategy consists of selecting only a percentage—say, the 75% of the dataset—and use the remaining 25% to test the algorithm. However, a small change in the dataset can cause a large change in the final predictions, and some adjustments are often required to increase the algorithm's robustness.

In the case of decision trees, it turned out that a tree that is grown very deep tends to suffer from high variance and low bias. This means that the tree is likely to overfit the training data: if we randomly split the training set into two parts, and fit a tree to both halves, the results could be quite different. To alleviate this lack of robustness, in the mid-1990s some data scientists argued that by aggregating many trees and perturbing the training set, using bagging (Breiman, 1996), boosting (Freund et al., 1996) or random forests (Ho, 1995), dramatically increased the predictive accuracy of the trees, by decreasing the variance. Bootstrap aggregating, or bagging, repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples. After training, predictions are averaged over the B samples: the method leads to better predictions because it decreases the variance of the model, without increasing the bias. Random forests improve over the bagged trees by using a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the predictors. The reason is that if one or a few predictors are very strong predictive for the response variable, these features will be selected in many of the B trees, causing them to become correlated. A new sample of approximately $\sqrt{p}$ predictors is taken at each split, by decorrelating the trees; predictions are then obtained as in bagging, by averaging over the B samples. Boosting works in a similar way to bagging and random forests, but the trees are grown sequentially, thus each tree is grown using information from the previous grown tree. This procedure does not rely on bootstrap sampling, instead each tree is fit on a modified version of the original dataset.

4. Predictive instrumentalism and how to make predictive models transparent and falsifiable

4.1. ML scientists are strong instrumentalist, statisticians are weak instrumentalist

As emerges from the quick overview of the well-known decision tree methods in the previous section, the only rationale to evaluate the goodness of an algorithmic modeling procedure is to look at its predictive accuracy on out-of-sample/future data. 'Shaking the training set' became popular to ensure lower variance and higher accuracy, with the data scientist apparently ready to do 'whatever it takes' to improve over the previous methods. From a philosophical and scientific point of view, algorithmic modelers are *strong instrumentalist*, since for them the predictive accuracy carried out by their algorithms is constitutive—and not only symptomatic—of the broader scientific success.

Evaluating a model/algorithm in light of its ability to predict future data is not shameful at all; conversely, it turned out to be beneficial in many areas where a parametric stochastic model failed to be really generative and useful. However, in line with Popper, predictions of future data are good tools to falsify a posed theory, but many times ML techniques lack of a general and valid theoretical 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:

$$p(\tilde{y}|y) = \frac{p(\tilde{y},y)}{p(y)} = \frac{1}{p(y)} \int p(\tilde{y},y,\theta)d\theta. \tag{3}$$

From Equation (3) we immediately notice that whenever we are interested in predictions, we need to consider a joint model $p(\tilde{y},y,\theta)$ for both the observed data y and the unobserved quantities $\tilde{y},\theta$. This joint model incorporates both the likelihood and the prior, being $p(\tilde{y},y,\theta) = p(\tilde{y}|\theta)p(y|\theta)p(\theta)$. Thus, the joint model for the predictions, the data and the parameters is transparently posed, and open to falsification when the observable $\tilde{y}$ becomes known.

5. Applied example: football Russia World Cup 2018

In this section we put in evidence the influence of the training set for future predictions by revealing some paradoxical considerations in ML results from a small-sample case. We consider here the dataset containing the results of all the 64 tournament's matches (48 of the group stages, and 16 of the knockout stage) for the FIFA World Cup 2018 hosted in Russia and won by France.

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:

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

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, $T = 1,\dots,N_T$; the nested indexes $h_n, a_n = 1,\dots,N_T$ 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

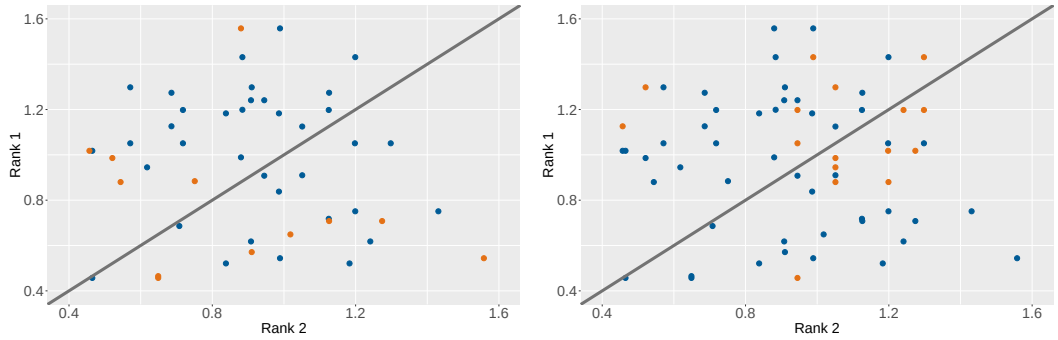
(b) *Train* Group stage matches
Test Knockout stage

(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.

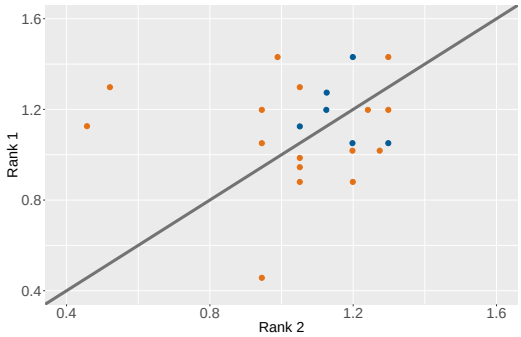
Table 1 shows the accuracy in the predictions for the seven methods and the three scenarios. As already argued, the choice of the training and the test set can dramatically change the predictive performance of the ML algorithms, which over-perform statistical models only when considering a portion of the group stage to predict the remaining group stage matches. Instead, it is worth noting that statistical models better predict the knockout stage matches (scenario B and C), and seem to be constant across the different scenarios.

Although the example is quite simple and the dataset is too small to extract general conclusions, there are enough arguments to emphasize the paradoxical performance achieved by



(a) Scenario A

(b) Scenario B



(c) Scenario C

Figure 1. For each prediction scenarios, the values of the FIFA rankings for each match are shown in blue colour for the training set and in orange colour for the test set.

Table 1. Prediction accuracy for the selected methods, according to three prediction scenarios.

Train	75% group	100% group	rank > 1
Test	25% group	knockout	knockout
Random forest	0.67	0.25	0.44
Bagged CART	0.67	0.31	0.37
CART	0.58	0.31	0.19
MARS	0.58	0.38	0.49
NN	0.67	0.25	0.44
Double Pois.	0.58	0.50	0.56
Biv. Pois.	0.58	0.56	0.56

ML techniques. Their predictive accuracy is too much influenced by the training set structure, making impossible to draw conclusions about their plausibility for predicting the football World Cup. In other way said, from this simple case study we cannot openly falsify our ML techniques on the ground of future predictions. Conversely, Poisson models are quite stable in the three scenarios in terms of predictive accuracy, and they seem to learn from the group stage training set in an efficient way. We are not claiming they are better than ML tools, we are just suggesting that they seem to be less sensitive to the training set structure, and then falsifiable in a broader sense.

6. Conclusions

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *In: Petrov, B.N., Csaki, F. (eds.) Proceedings of the Second International Symposium on Information Theory, pp. 267281. Akademiai Kiado, Budapest (1973). Reprinted in: Breakthroughs in Statistics*, pp. 610–624. Springer, New York (1992).

Baio, G. and M. Blangiardo (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics 37*(2), 253–264.

Breiman, L. (1996). Bagging predictors. *Machine learning 24*(2), 123–140.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science 16*(3), 199–231.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and regression trees*. Wadsworth.

Bzdok, D., N. Altman, and M. Krzywinski (2018). Points of significance: statistics versus machine learning.

Freund, Y., R. E. Schapire, et al. (1996). Experiments with a new boosting algorithm. In *icml*, Volume 96, pp. 148–156. Citeseer.

Friedman, M. (1953). *Essays in positive economics*. University of Chicago Press.

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.

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.

Vehtari, A., J. Gabry, Y. Yao, and A. Gelman (2019). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.1.0.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing 27*(5), 1413–1432.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research 11*(Dec), 3571–3594.

Worrall, J. (2014). Prediction and accommodation revisited. *Studies in History and Philosophy of Science Part A 45*, 54–61.