Weapons of mass prediction

Leonardo Egidi[a] (joint work with Jonah Gabry[b], in preparation for *Journal of Royal Statistical Society, Series A)*

legidi@units.it

November 22nd, 2019

StaTalk 2019

[a]Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche *Bruno de Finetti*, Università degli Studi di Trieste, Trieste, Italy

[b] Department of Statistics, Columbia University, New York, USA

Outline

The role of prediction in science

Weapons of mass prediction

Weak instrumentalism

Some examples from my/our research

References

The role of prediction in science

The Delphi's Oracle

The role of prediction in science

- Falsificationist philosophy of Karl Popper [Popper, 1934]: theories, in order to be scientific, must be **falsifiable** on the ground of their predictions.
- Wrong predictions should push the scientists to reject their theories or to re-formulate them, conversely exact predictions should corroborate a scientific theory.
- Strong instrumentalism [Hitchcock and Sober, 2004]: predictive accuracy is constitutive of scientific success, not only symptomatic of it, and prediction works as a **confirmation theory** tool for science.

The role of prediction in (data) science

- **20th century**: expansion of science's boundaries. Not only psysics and natural science, but social and computational sciences as well.
- Probabilistic and statistical methods have made the 'debut of science in society' possible.
  - **1940's**: Manhattan Project in Los Alamos, MCMC techniques (Enrico Fermi, John Von Neumann, Stanislaw Ulam).
  - **1970's**: GLMs (McCoullagh, Weddenburn)
  - **1980's**: Neural Nets, Decision Trees. R
  - **1990's**: WinBUGS, automatic MCMC procedures.
  - **2000's**: Random Forests, Machine Learning
  - **2010's**: Stan, Deep Learning
- **Main question**: are social sciences falsifiable in light of their predictions? Is a theory/model good only if able to well predict future events?

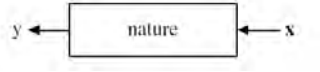
When falsification does not make sense: Greece, Leicester, Trump, Brexit...

Weapons of mass prediction

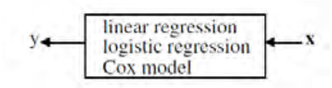
Statistics and Machine Learning

Statistics and Machine Learning

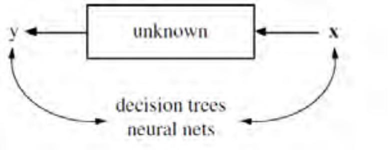
- **Two cultures** [Breiman et al., 2001]: link between some input/independent data x and some response/dependent variables y.
- Nature: unknown



- Statistics : information



- Machine Learning: prediction



Weapons of mass prediction

- Statistics and Machine Learning: most popular 'prediction's weapons' for social and natural sciences (weather forecasting, Presidential elections, global warming, etc.).
- Though, many times the right weapons are embraced by the wrong people.
- The predictive power in statistics is an elegant, small **gun**, with good properties but small bullets, whereas in machine learning is a **bazooka**, with devastating effectiveness and big bullets.
- Usually, statisticians do not take into account predictions as confirmation tools for their theories, conversely Machine Learners care predictions too much. Maybe, we need something in between.

Predictive model's accuracy in statistics

- **Predictions' uncertainty**: in our practice, prediction should not be assimilated to 'take a rabbit out of a hat', but looking at its inherent uncertainty.
- **Posterior predictive distribution**: future hypothetical values $\tilde{y}$ come from a probability distribution, $p(\tilde{y}|y)$, such that we could define an expected predictive density (EPD) measure for a new dataset.
- **Predictive information criteria**: Watanabe-Akaike Information Criteria (WAIC) [Watanabe, 2010] and Leave-One-Out cross validation Information Criteria (LOOIC) [Vehtari et al., 2017]: data granularity, by definition of the log-pointwise predictive density $p(\tilde{y}_i|y)$ for each new observable value $\tilde{y}_i$.

Predictive accuracy in Machine Learning

- **Training set** choice: select the first half, or a percentage of a dataset to train the algorithm, and use the remaining portion to test the algorithm.
- **Lack of robustness**: 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.
- **Overfitting**: a decision tree that is grown very deep tends to suffer from high variance and low bias, 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: Random Forests, Boosting, Bagging.

Weak instrumentalism

Maybe not too weak...

Weak and strong instrumentalism

- **Statistics**: predictions and predictive accuracy are only sometimes constitutive of scientific success (weak instrumentalism). Usually, the only rationale to evaluate the goodness of a statistical model is to look at its residuals. *We need something more!*
- **Machine Learning**: predictive accuracy on out-of-sample/future data is the only rationale to evaluate the goodness of ML procedures (strong instrumentalism). *We do not need just this!*
- **Goal**: produce good, transparent and well posed algorithms/models, and make them falsifiable upon a strong check [Gelman and Shalizi, 2013].

Falsificationist Bayesianism: beyond inference and prediction

- **Falsificationist Bayesianism**: model checking through pp checks. Prior: testable part of the Bayesian model, open to falsification [Gelman and Hennig, 2017].
- $\tilde{y}$: unobserved future values, with posterior predictive distribution:

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

where $p(\theta|y)$ is the posterior distribution for θ, whereas $p(\tilde{y}|\theta)$ is the likelihood function for future observable values. Equation (1) may be resambled in the following way:

$$p(\tilde{y}|y) = \frac{p(\tilde{y}, y)}{p(y)} = \frac{1}{p(y)} \int p(\tilde{y}, y, \theta)d\theta. \tag{2}$$

A joint model $p(\tilde{y}, y, \theta)$ for the predictions, the data and the parameters is transparently posed, and open to falsification when the observable $\tilde{y}$ becomes known.

Limits of Machine Learning predictions

- **Tuning parameters**: the number of predictors at each split of a random forest is a tuning parameter fixed at $\sqrt{p}$ in most cases, but in practice the best values for these parameters will depend on the problem.
- **'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.
- **Generalization**: how well the concepts learned by a machine learning model apply to specific examples not seen by the model when it was learning. Ideally, you want to select a model at the sweet spot between underfitting and overfitting. This is the goal, but is very difficult to do in practice!

So, what is weak instrumentalism, actually?

- **Transparency**: 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.
- **Pre-existence**: supposedly valid scientific theories should exist *before* the future data have been revealed, and produce some immediate benefits to the scientific community.
- Weak instrumentalism's main task is to make statistics more predictive (e.g., using a joint model for predictions, data and parameters, as in falsificationist Bayes) and Machine Learning more explicative.

Summary table

Table 1. Weak instrumentalism summary

General science

p1 Predictive accuracy is not always constitutive of scientific success
p2 Scientific falsification on the ground of wrong predictions is sometimes misleading, especially in social sciences (Trump's election, Leicester win, Brexit)
p3 Supposedly valid scientific theories should exist before the future data have been revealed
p4 Prediction is not explicitly part of the formulation of a scientific hypothesis at the time the law is posed, but it becomes relevant and relevant as science advances

Statistics

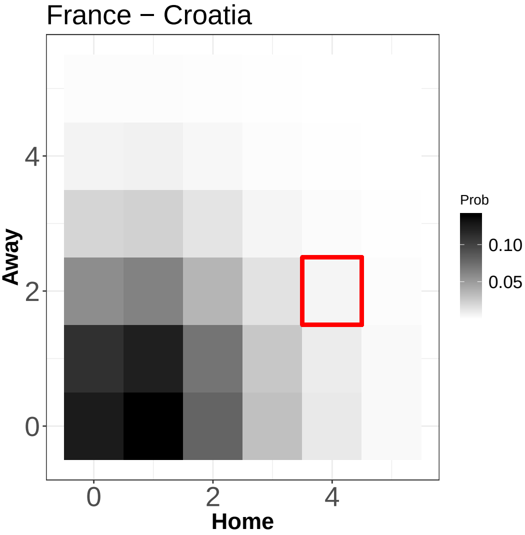
p5 Take care of variability in the statistical predictions
p6 If necessary, go beyond the distinction between inference and prediction, and consider a joint model for data, parameters and future data (falsificationist Bayes)
p7 Rather than reasoning in terms of variance and bias, reason more in terms of predictive information criteria and posterior predictive distribution

Machine Learning

p8 'Shaking the training set' to improve predictive accuracy is an obscure step
p9 Avoid to tune the algorithm with the only task to improve predictive accuracy
p10 To be falsifiable, ML techniques need to be transparently posed

Some examples from my/our research

Posterior probabilities for the World Cup 2018 final



footBayes R package

(available at: https://github.com/LeoEgidi/footBayes)

Accuracy for World Cup predictions

Table 2. 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

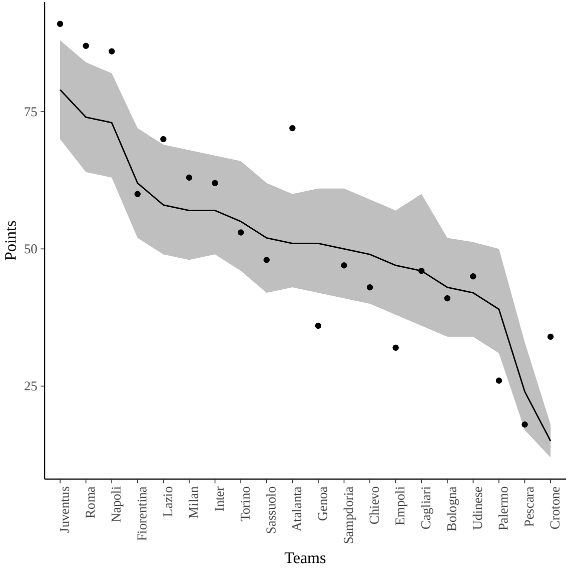
[Egidi and Torelli, 2019]

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.

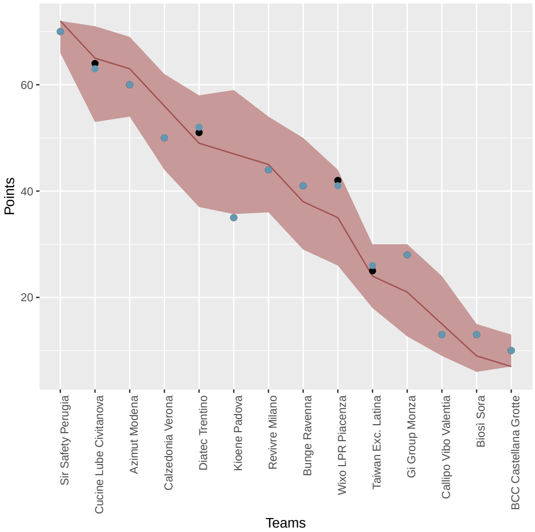
Prediction of the final rank league: Serie A 2016-2017



Predictive intervals for the final number of points

Egidi et al. [2018]

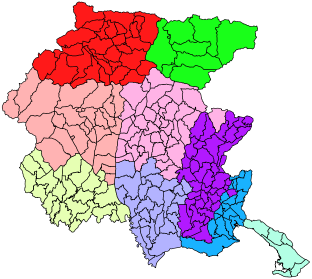
Prediction of the volleyball rank: SuperLega 2017-2018

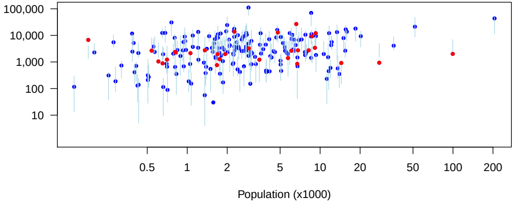


Predictive intervals for the final number of points

Egidi and Ntzoufras [2019]

Prediction for the FVG commuters





Confidence bars for the number of FVG commuters

Egidi et al. [2019]

Discussion

- Prediction and predictive accuracy are central in the progress of science and became even more relevant in statistics and data science.
- Though, social sciences are not falsifiable as physics and natural sciences.
- As statisticians demanded to build **good models** to accomodate complex data, we feel that predictive accuracy is not always constitutive of scientific success: prediction is not everything, however is vidal, and it is our responsibility to choose between the gun or the bazooka.
- **Weak instrumentalism** philosophical view is designed to alleviate the falsification issue raised by strong instrumentalism and to provide a bunch of rules to make Statistics and Machine Learning more transparent.

Put Statistics and ML far from these guys!



References

References

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

Leonardo Egidi and Iannis Ntzoufras. A Bayesian quest for finding a unified model for predicting volleyball games. *Submitted to Journal of Royal Statistical Society Series C (Applied Statistics)*, 2019.

Leonardo Egidi and Nicola Torelli. Comparing statistical models and machine learning algorithms in predicting football outcomes. *Conference paper, Statistics for Health and Well-being, 2019, Book of Short Papers*, 2019.

References ii

Leonardo Egidi, Francesco Pauli, and Nicola Torelli. Combining historical data and bookmakers' odds in modelling football scores. *Statistical Modelling*, 18(5-6):436–459, 2018.

Leonardo Egidi, Francesco Pauli, Nicola Torelli, and Susanna Zaccarin. Clustering spatial networks through latent mixture models. *Under review in Advances in Data Analysis and Classification*, 2019.

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

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

References iii

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

Karl Popper. *The logic of scientific discovery*. Routledge, 1934.

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

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