Prediction is not everything, but everything is prediction

Abstract

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

Keywords: Prediction; Popper?s falsification philosophy; Weak instrumentalism; Predictive accuracy; Machine learning

1 Introduction

The ultimate goal of science is doubtless to find the *truth* while testing theories, and this sparked a lot of debate about two distinct positions in the philosophy of science, namely *realism* and *empiricism* (Sober, 2002). According to the former, the main scientific goal is to test and check which theories are true, whereas the latter suggests to discover which theories could be empirically suited for the problem at hand. In both cases, truth is the primary focus. According to another philosophical position, *instrumentalism*, science should instead provide accurate predictions rather than assessing whether a theory is true or not: in other words, the search for true theories and that for accurate predictive theories coincide; on the other hand, many philosophers of science claim that prediction has a primary role in the progress of science and, accordingly, how predictive accuracy is one of the ultimate goals in scientific inference (Forster, 2002). The interesting interplay existing between truth and prediction accuracy invokes *falsificationism* (Popper, 1934) with related criticisms (Kuhn, 1962; Lakatos, 1976). Popper argues in fact that theories, to be scientific, must be falsifiable on the ground of their predictions: wrong predictions should perhaps push scientists to reject their theories or to re-formulate them, conversely exact predictions should corroborate a scientific theory. Using the above terminology, we could say that 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.

With regard to statistics, we are continuously pushed and asked to build, develop and test scientific theories by using a bag of tools containing distinct items, such as hypothesis testing, likelihood theory, Bayesian methods, randomized trials, non-parametric methods,

and so on. As statisticians, we daily construct elegant devices (models) in many scientific disciplines, such as economics, psychology, physics, biology, education, and environmental sciences, designed to: (i) formulate a theory (ii) test the theory from some evidence (data) (iii) generalize the theory by induction (iv) confirm the theory. By adopting the above terminology, as statisticians we feel to belong to the tradition of empiricism, in line with the well-known George E. P. Box's quote *"All models are wrong, but some are useful"*: we could then never 'sell' our models as *true* and real theories, rather we could just assess their empirical adequacy. However, this assessment is nowadays the matter of a vivid debate within the statistical community about the distinction between explanatory and predictive modeling (Shmueli, 2010; Hitchcock and Sober, 2004), particularly when the two *dimensions*—explanation and predictions—result to be in conflict with each other. To clarify, we define here explanation or explanatory modeling the use of statistical models for testing causal hypotheses or explanations—e.g. between a set of covariates and a response variable; we define prediction or predictive modeling the action of using a model (or device, algorithm) to produce predictions for new or future observations. The milestone of this debate can be considered as one of the most influential papers ever, Breiman et al. (2001), where Leo Breiman shed light on two distinct cultures, the modeling and the algorithmic one, which usually raised the importance of causal explanation and prediction, respectively.

However, these two dimensions often appear conflated, and any attempt to state the supremacy of one of the two entities over the other is in most cases misleading. For such reason, we think the debate about the supposed prevalence of one of the two dimensions is local and application-dependent and, in general, wasteful. Rather, we could try to go beyond, by incorporating prediction in the explanation step. As statisticians, we spent most of our efforts and our best years for reporting statistical estimates and standard errors,

along with measures of *statistical significance* (Gelman, 2016d), for *unobservable* parameters that do not exist in real life; rather, in this paper we maintain with Billheimer (2019) that inferential statistics is *inherently predictive* and should be focused on probabilistic predictions of future *observable* events and quantities. In such a way, the predictive inference paradigm (Bjornstad, 1990; Geisser, 2017), somehow in line with the prequentialism theory of Dawid (1984), allows the comparison of competing scientific theories by way of predictive accuracy and avoids to emphasize the estimation of virtual and artificial parameters. Moreover, from a communication point of view observed values are easier to interpret and much more accessible to non-statisticians than artificial parameters.

However, taking prediction to assess and explain model capability does not solve all the problems surrounding the statistical discipline. Nowadays, statistical models are currently used with application to very different disciplines, ranging from the so-called 'hard sciences' such as physics, biology and engineering to some social sciences such as psychology, education and economics. In this sense, we do not want to be entirely instrumentalist: rather, in this paper we encourage the use of predictive inference and propose the *weak instrumentalism* (herafter, WI), under which predictive accuracy is constitutive of scientific success only when the underlying statistical methods are falsifiable and transparently designed to predict future and new events. In other words, there are many contexts, especially in the social sciences, where falsification through the prediction's fallacy should be replaced by a more consistent idea of falsification: we believe this position may be beneficial for the so-called hard sciences as well. On one hand, mathematical and quantitative laws formulated by Galilei and Newton were physical and deterministic with which particular future facts could have been predicted with absolute precision; on the other hand, probabilistic and statistical laws designed to describe human behavior and social facts are stochastic laws,

with which particular future events could be predicted with an intrinsic amount of uncertainty. Of course, as statisticians we want to do our best in predict future social events, but we cannot entirely evaluate a model's performance only on the ground of its predictive accuracy. Using Popper's terminology, incorrect social sciences predictions should not be the only tool to falsify a theory.

From the arguments above it emerges clearly how prediction is relevant and instrumental for scientists working with data: however, it is not all we need, especially when framed in social science frameworks. Perhaps we need transparent assumptions, acknowledgment of variability in the distribution of the observables, considerations about the legitimation of the inductive reasoning. And it is clear how this legacy, in the modern big-data age with vast amounts of available data, should involve the whole data-science field, not only the theoretical statisticians or the *data-miners*: the WI philosophy should push model developers to criticize themselves and to eventually reformulate their algorithms in order to improve the scientific inference.

The remainder of this paper is organized as follows.

2 The role of prediction in hard and social sciences: a brief overview

In a very popular and well-known essay, Russell (1931) summarized the three main stages required to formulate a scientific law as follows: (1) observation of some relevant facts, (2) formulation of a hypothesis underlying and explaining the previously mentioned facts, and (3) deduction of some consequences from this hypothesis. As suggested by the same Russell, the modern scientific method was born with Galileo Galilei, father of the law

of falling bodies, and with Johannes Kepler, who discovered the three laws of planetary motion (from Russell (1931)):

> *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.*

It is easy to see how the sequence of events that occurred over the next 250 years told us much about how scientific inference proceeds, with the law of universal gravitation of Isaac Newton embodying the two previous theories (18th century), and the theory of the general relativity of Albert Einstein (20th century) generalizing itself Newton's theory. It is then clear how in the last 500 years, physics—and, more generally, science—advanced by falsification and generalization of previous theories, by providing new and more exciting theories to predict new natural facts and highlighting the confirmation nature of prediction. Perhaps, 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 role of prediction with regard to scientific progress is more ambiguous than what people are usually inclined to think. In general, is prediction a central step in science, or even, in the words of Forster (2002), is prediction the ultimate goal of science, according to the instrumentalist position (Sober, 2002; Hitchcock and Sober, 2004)? From an instrumental perspective, predictive success is not merely *symptomatic* of scientific success, is also *constitutive* of scientific success (Hitchcock and Sober, 2004). On the other hand, for Popper (1934), a theory is scientific only if it is falsifiable, where the falsification

of a theory is meant to be the possibility of comparing 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.

To complicate the debate, over the last decades scientific predictions has become popular not only in the context of physics and natural science, but for the social sciences as well. As argued by Popper (1944, 1945) and Sarewitz and Pielke Jr (1999), the social sciences have long tried to emulate physical sciences in developing invariant mathematical laws of human behavior and interaction to predict economics quantities, elections, policies, etc., being the modern data-scientists increasingly asked to build *'weapons of mass prediction'* in varying social contexts. Even though many scholars agree that a social theory should be judged on its power to predict (Friedman, 1953), we feel however that an instrumentalist position about predictive accuracy is still ambiguous in the social sciences, especially when the final actual observed outcome is somehow far away from the model's predictions. To take some explanatory and well-known examples, consider some very low-probability events, all actually occurred during 2016, such as the Donald Trump's win against Hillary Clinton in the US presidential elections, UK Brexit, and the Leicester's Premier League's win in soccer. None of the best data-scientists and modelers would bet even a dollar on each of these events happening—consult Gelman (2016a,b,c) for some more considerations about these facts. Can all of these rare, though actually observed, events falsify the finest algorithms and models that did not correctly predict their occurrence? Does it make sense to criticize such social science models/algorithms by retrospective evaluations of probabilistic predictions? We believe that social science methods require more valid motivations to validate or falsify underlying theories beyond that of providing *bad* predictions for rare

events (Gelman et al., 1998), such as a deep inspection of quality of data, psychological attitudes, study design, inclusion of more/other covariates and eventual interactions, and so on. Scientific falsification detected by wrong predictions (Popper, 1934) is a powerful and exceptional tool, but in this paper we caution its abuse/misuse. We give more details on this weak instrumentalist position in the next sections.

3 The role of prediction in statistical learning

3.1 From the observed to the observable

As frequentist or Bayesian statisticians, we often deal with a double task: first, creating a sound mathematical model to accommodate the data and retrieve useful inferential conclusions from parameters' estimates; second, using this model to make predictions for future and new observations. However, statistical theories are usually tested and built by way of causal *explanation* and parameters' estimation, whereas predictive modeling is usually considered much less important—if not irrelevant and unscientific—to develop scientific theories (Shmueli, 2010): perhaps, formulate a theory for an unknown mechanism by *explaining a theoretical causal relationship* is often considered the dominant side and seen as separate from prediction, which instead moves from the observed to the unobserved (though observable in the future), being the action designed to forecast future or new events without requiring a full understanding of the underlying data-generation process. Predictive actions can be daily shared and accepted from the human common sense: each person is in fact more or less confident with 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, we got

a paradoxical argument, since inference, or explanation, seems difficult and obscure, and prediction simple and transparent to the non-statisticians. Statisticians contributed much to feed this paradox. In the past they operated like those magicians who resemble their decks of cards by fictitiously adding/removing extra cards that are not available to the audience. To continue with the metaphor, model parameters behave like these fake cards, which are incredibly relevant to build the trick (aka statistical theories), but do not exist in the real life; in brief, *parameters are fictitious and technical devices used to explain and approximate complexity*. Rather, only prediction links the observed with the observable and is accessible to the people: it is never a matter of parameters' interpretation (Billheimer, 2019), it only requires a check of the discrepancy between observed and future events, and can doubtless be done by anyone.

We just need to admit that, when framed in a predictive task, we feel a bit lost in the space, deprived of our comfortable tools, and not ready to deal with most of the technical questions and difficulties that arise: the choice of the training and the test set, the over-fitting phenomenon, the measurement of predictive accuracy, the use of information criteria, and so on. We should instead accept prediction in our practice and possibly 're-brand' our inferential mechanism by investing much more on prediction. For illustrations purposes, consider a logistic/probit regression model for diagnosing diabetes testing positive probabilities using biochemical predictors such as glucose, insulin, mass, etc.. A summary for these kinds of models is usually documented by the adoption of odds-ratios, confidence/predictive intervals for the parameters, p-values and other estimation-driven measures. However, the hidden task of such model is inherently predictive: rather than over-focusing on the numerical impact of the frequentist/Bayesian estimates, the focus here should be on the future probability of being tested positive to the diabetes, by considering

existing or new values for glucose, insulin, etc.. As another prototypical example, consider the illustration proposed by (Billheimer, 2019, Section 4), where a trial consisting of two treatment groups of ten mices each is observed. Rather than focusing on inferential conclusions from hypothesis testing, the author carries out a predictive inference application by producing the predictive distribution of replicated and observable experiments and reporting the uncertainty surrounding the difference in the sample means. In general, similar clinical trial studies are rarely conducted with the idea of predicting a future event for the population of interest: rather, they are usually built to find and display a significant effect, according to the statistical significance of some parameters. This estimation obsession makes statistics obscure, whereas predictive inference, openly accessible even to non-statisticians, would make statistics more transparent and falsifiable. This is the reason why we provocatively claim that *prediction is not everything, but everything, in statistics, could be described in predictive terms.*

3.2 Good and bad practices

We need to reason whether the statistical predictive tools are useful for the *desiderata* we stated in the previous section. If we wish to make prediction a reliable tool to explain the development of some theories, we should guarantee that the usual tools are appropriate to this task. In short, how we asses and measure predictive accuracy? Moreover, is this assessment worth to provide useful explanations about our theory?

A typical assessment is based on information criteria, such as Akaike Information Criterion (AIC) (Akaike, 1973), Bayesian Information Criterion (BIC) (Schwarz, 1978), and Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). Though, the three metrics above exhibit two main flaws: they do not provide any kind of prediction uncertainty

and they rely on the number of *nominal* parameters included in the model (Gelman et al., 2014): for these reasons, their use should be limited to a restricted case of applications. 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), naturally framed in a Bayesian paradigm, incorporate the uncertainty propagating from the parameters to the distribution of observable values by estimating the expected log pointwise predictive density for a new dataset. Although all the predictive information criteria may fail in some practical situations, LOOIC and WAIC acknowledge the intrinsic model uncertainty, by highlighting the role of the observable values, rather than parameters. A transparent predictive and explainable tool should encompass data, parameters and future data not focusing on parameters plug-in estimates alone; in such a way, the falsification of a single piece makes the joint model falsifiable.

Another typical assessment is given by the generalization performance of a learning method related to the predictive performance on a test set; unfortunately, it is commonly not feasible to calculate the test mean-squared errors, because we are often in a situation in which we do not have any test data available. Hastie et al. (2009) propose to use the *expected test MSE*, or *expected generalization error*, usually estimated by cross-validation and related statistical methods, and resulting as the sum between the variance, the squared bias, and the irreducible error of the posed algorithm. The predictive goal is then to select the model where the expected test MSE is lowest, by choosing the model that simultaneously has low variance and low bias. It is in fact well-known that the performance of a statistical learning method requires low variance as well low bias: the challenge is to find a compromise by controlling for both bias and variance. However, this is a relevant and difficult challenge, widely known as the *bias-variance trade-off* (Hastie et al., 2009; James et al., 2013), much

debated within the statistical community: 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.

Concerning the bias-variance dilemma, we think this is usually seen as the *Holy Grail* of the modern statistical learning, and we would try to slightly falsifying its importance by highlighting at least four criticisms for this predictive protocol.

(i) **Applicability** It is pretty impossible to monitor and assess bias and variance of a statistical learning method in real-life applications.

(ii) **Prescription** Even though we were able to compute them and retrieve a valid assessment for the expected test MSE, we would not have neither a scientific prescription about the predictive accuracy and the potential overfitting of our algorithm: *how and when does a model provide overfitting? Do we have a numerical, or somehow objective, scale to assess this task?*

(iii) **Data-treatment** Over the last years, statisticians developed some methods by having in mind to reduce their variance and improve their predictive accuracy, by aggregating thousand versions of the same method by use of bootstrap (Breiman, 1996), eventually perturbing the training set (Freund et al., 1996) and the choice of the predictors to ensure de-correlation (Ho, 1995) in classification and regression trees (Breiman et al., 1984). This pushed the statisticians to do *whatever it takes* to produce effective *weapons of mass prediction*, including bad practices such as *shaking the training set* in order to obtain the best prediction, or suddenly change the number of predictors and the correspondent tuning parameters used in a given random forest split, or the interaction depth parameter in boosting methods (*over-tuning*). Modern data-scientists are used to train their procedures on the *training set*, which is chosen

at the beginning. 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,75% of the dataset—and using 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.

(iv) **Interpretability, transparency and falsification** Points (i)–(iii) above contributed to build very powerful, complex, and obscure black-box predictive algorithms, such as neural networks (Bishop, 1994), which are rarely interpretable, explainable (Xu et al., 2019) and falsifiable from a scientific point of view. *How could we fruitfully use the same method in other and alternative settings if we do not know how we built it?*

3.3 Weak instrumentalism for statistical learning

As statisticians, demanded to build models for social and physical sciences, our efforts should be addressed to produce good, transparent and well posed statistical learning methods, and possibly make them falsifiable upon a strong check of their components (Gelman and Shalizi, 2013). However, as mentioned in the previous section, the goodness of a modeling procedure is often associated with its predictive ability on out-of-sample scenarios. As a consequence, only good predictive models are retained, whereas the others, even when sophisticated and well built, are discarded; predictive accuracy became in many fields the only tool to decide between good and bad statistical methods. Points (i)–(iv) in Section 3.2 refer to the *strong instrumentalism* philosophical position, for which the predictive accuracy carried out by the algorithms is constitutive—and not only symptomatic—of broader scientific success.

Even though it is worth stressing that 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, for instance where a parametric stochastic model failed to be really *generative* and useful. However, even if predictions of future data were good tools to falsify a posed theory, some strong instrumentalist techniques lack a general and valid theoretical framework. As an illustrative example, the number of predictors at each split of a random forest is a tuning parameter fixed at the square root or even at one third of the number of total predictors in most cases, but in practice the best values for these parameters will depend on the problem at hand; if the final method is tuned and selected on the ground of its predictive accuracy, the underlying theory to be falsified is artificial, and not posed in a transparent way.

It is clear to us that a valid contribution to a transparent and falsifiable science should embrace both methods' explainability and predictive accuracy. To do this, we feel we are ready to propose a *weak instrumentalist* philosophical position, a sort of behavioral protocol, that can be listed in the following bullets.

(1) **Applicability** Social sciences are different from hard sciences. Then, statistical methods for social sciences should not be discarded just because of their poor predictive accuracy.

(2) **Prescription** predictive inference (Billheimer, 2019); Distribution of the observables, falsificationist Bayesianism (Gelman and Hennig, 2017);

(3) **Data-treatment** not invasive wrt predictive accuracy results; sensitivity analysis fully reported.

(4) **Interpretability, transparency, and falsification** joint model for parameters,

observed and observable. Variability. Predictive distribution for NHST. Explanable AI

arrivato qui. Fare una tabellina di confronto tra strong e weak instrumentalism.

4 Discussion

Da scrivere ex-novo.

SUPPLEMENTARY MATERIAL

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).

Billheimer, D. (2019). Predictive inference and scientific reproducibility. *The American Statistician 73*(sup1), 291–295.

Bishop, C. M. (1994). Neural networks and their applications. *Review of scientific instruments 65*(6), 1803–1832.

Bjornstad, J. F. (1990). Predictive likelihood: A review. *Statistical Science*, 242–254.

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.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General) 147*(2), 278–290.

Forster, M. R. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of science 69*(S3), S124–S134.

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.

Geisser, S. (2017). *Predictive inference: an introduction*. Chapman and Hall/CRC.

Gelman, A. (2016a).

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

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

Gelman, A. (2016d). The problems with p-values are not just with p-values. *The American Statistician 70*(10).

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., J. Hwang, and A. Vehtari (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing 24*(6), 997–1016.

Gelman, A., G. King, and W. J. Boscardin (1998). Estimating the probability of events that have never occurred: when is your vote decisive? *Journal of the American Statistical Association 93*(441), 1–9.

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.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

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.

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.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Shmueli, G. (2010). To explain or to predict? *Statistical science 25*(3), 289–310.

Sober, E. (2002). Instrumentalism, parsimony, and the akaike framework. *Philosophy of Science 69*(S3), S112–S123.

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.

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.

Xu, F., H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu (2019). Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pp. 563–574. Springer.