International Journal of Educational Research 87 (2018) 36–46

Contents lists available at ScienceDirect

International Journal of Educational Research

journal homepage: www.elsevier.com/locate/ijedures

The power of noise and the art of prediction

ZhiMin Xiao*, Steve Higgins

School of Education, Durham University, United Kingdom

ARTICLE INFO

Keywords:
Cross-validation
Evidence-based policy
K-NN
Logistic regression
Prediction
Random forests

ABSTRACT

Data analysis usually aims to identify a particular signal, such as an intervention effect. Conventional analyses often assume a specific data generation process, which implies a theoretical model that best fits the data. Machine learning techniques do not make such an assumption. In fact, they encourage multiple models to compete on the same data. Applying logistic regression and machine learning algorithms to real and simulated datasets with different features of noise and signal, we demonstrate that no single model dominates others under all circumstances. By showing when different models shine or struggle, we argue that it is important to conduct predictive analyses using cross-validation for better evidence that informs decision making.

1. Two modelling approaches

Data analysis is usually about identifying signal from noise. But data, particularly social science data, can be truly noisy, partly because the outcome is often a human construct, which can only be measured with some error. Noise can also stem from other factors, such as the collection of data on variables that are not correlated with the outcome of interest, or the addition of interaction and/or higher order terms, which can easily fail in-sample goodness-of-fit tests (Breiman, 2001), meaning the combination of many variables and their various transformations in a conventional, say, linear regression, can be so flexible or fit the observed data so well that it has little value in the explanation of new or out-of-sample data. The noise mentioned above can be minimised in careful research designs and sound data analyses. Nevertheless, theories about best designs and views on best analysis strategies can be another source of noise, because the best approach to data analysis for a given study often differs in theory from person to person, even for those who are from the same discipline (Xiao, Kasim, & Higgins, 2016). One classical example is a statistical phenomenon called Lord's Paradox (Lord, 1967, 1969), where the relationship between two variables change in both magnitude and direction when a third variable is statistically controlled for (Holland, 2005; Tu, Baelum, & Gilthorpe, 2008; Tu, Gunnell, & Gilthorpe, 2008; Wainer, 1991; Wainer & Brown, 2004; Xiao, Higgins, & Kasim, 2017a). In education, decisions on not just variables but also types of models are often necessary, in order to account for the fact that pupils are usually nested within classes, which are from schools located in different regions. Depending on the structure of a specific dataset, such choices can result in considerable differences in point estimates and uncertainties surrounding those estimates (Xiao et al., 2016; Xiao et al., 2017a; Xiao, Higgins & Kasim, 2017b).

Moreover, single best models, or the practices of using just one model and explaining why that model is the best according to a mathematical theory or evidence found elsewhere, cannot be statistically compared unless some of them are nested within others (Shmueli & Koppius, 2011). Donoho called the analytical approach that relies on a single best model derived from a mathematical formula "generative modeling" (2015), where a data generation process is assumed and a single best model, which must exist because of the assumptions made, is then deployed to analyse the data. But the choice of that model can itself be a source of variation in results because of the theoretical differences mentioned above. Consequently, the model may lead to "irrelevant theory and

* Corresponding author at: School of Education, Durham University, Durham, DH1 1TA, United Kingdom.
E-mail address: zhimin.xiao@durham.ac.uk (Z. Xiao).

https://doi.org/10.1016/j.ijer.2017.10.006
Received 8 August 2017; Received in revised form 13 October 2017; Accepted 25 October 2017

questionable scientific conclusions" (Breiman, 2001) because it is usually more about a data generation and selection process than about how the real world functions or the underlying problem to be solved. When published, the results may further justify the choice of the theoretically best model in subsequent studies, particularly when they are linked with research funding streams, which in turn make the results more salient or more noticeable in the literature. This feedback loop can be pernicious (O'Neil, 2016), if policy decisions made on the evidence from a single best model produce unintentional and undesirable consequences (Merton, 1936) to those who participated in the studies and/or beyond.

Most regression models we see in educational research are of the generative type, which can be theoretically best because of the asymptotic guarantee, meaning if an intervention were to be repeated many times until all samples in a population are exhausted, the model is guaranteed to predict the correct outcome. This sounds reassuring, but in reality, we do not live in an "asymptopia" (Domingos, 2012), an imaginary situation where an intervention can be replicated many times without any constraint. This implies that, if model A is better than model B given infinite data, due to bias-variance trade-off or balance between precision and uncertainty, there is no guarantee that the former will be better than the latter given finite data or a particular dataset. As such, we also need "predictive modeling" (Donoho, 2015; Hofman, Sharma, & Watts, 2017), which is generally agnostic about a data generating mechanism and focuses on how well it predicts the future rather than how well it fits the data after the fact (Popkin, 2015). Typically, predictive modelling encourages multiple models to learn from and work on multiple datasets, some of which are used to train the models, others put aside as test sets, just as we turn a ball many times and each time we make a prediction about the patterns on the side we do not see using the information on the side we can see. The performances of the trained models are then judged against a common task, usually, predictive accuracy on test sets, which is easy-to-understand and can be compared across datasets and over time (Breiman, 2001; Donoho, 2015; Hofman et al., 2017; James, Witten, Hastie, & Tibshirani, 2015).

In education, when predictions are made on learning outcomes, it has the spirit of predictive modelling, which relies on some observed data in the past (training sets) and its performance is assessed on how accurately it predicts yet-to-be-observed outcomes (test sets). Note that predictive models do not have to be sophisticated. Assuming we know nothing or little about the past, guessing is one predictive model, which is usually less accurate than the averaging of past outcomes as another predictive model. As we gather more data, we can employ more powerful models such as regression to make more accurate predictions. Predictive modelling thus embraces changes in the real world and always improves as we feed it with more data (Popkin, 2015).

Predictive modelling also provides timely feedback for analysts to assess how successful the tools they have deployed actually are in the wild. As a result, the best performing models can be efficiently utilised for real-world applications, which can then enhance the roles evidence has to play in decision making and reduce the gap between research and practice (Shmueli, 2010). This approach overcomes one problem of single best models, where different analysts analyse the same dataset in their own manner and may produce different results and make different claims about the performance of their preferred methods (Breiman, 2001; Donoho, 2015; Hand, 2006; Xiao et al., 2016). If we do not know which, if any, of the best models actually worked because of the problems associated with in-sample strength-of-fit measures, such as the coefficient of determination or R^2 in a linear regression (Breiman, 2001; Shmueli & Koppius, 2011), the conclusions drawn from the results of single best models may be too dependent on error or noise, making them effectively just noises themselves. This only adds to the challenge of evidence-informed policy and practice by confusing decision makers with varying advice.

In many social science studies, such as the educational interventions funded by the Education Endowment Foundation (EEF) in the UK, predictive modelling is yet to be widely appreciated (Shmueli & Koppius, 2011), despite the aforementioned advantages and its rapid development and application in other fields (DeRubeis et al., 2014; Hofman et al., 2017; Kapelner et al., 2014; Kennedy, Wojcik, & Lazer, 2017; Popkin, 2015; Tetlock, Mellers, & Scoblic, 2017), such as the engineering of personal computing and smartphones in our daily lives, and individualised or precision medicine in scientific research. For the time being, the design and analysis of EEF trials focus primarily on average treatment effect on the treated, which is helpful if we are only interested in the mean. But one technical difficulty of the approach is that many large-scale EEF trials are producing negligibly small effect sizes, which has many implications, including an ethical one of randomly assigning students to a group that we could have predicted to be non-optimal for specific subgroups of students, given the evidence gathered earlier about a trial and the data collected about the students under concern. Also, when an intervention does not specifically target a subgroup, such as Free School Meal (FSM) pupils in England, an estimate of effect is always reported for the group with caution, which is analogous to saying: this intervention has such an effect on FSM pupils, but the public should not really trust the result. Subgroup analysis is notoriously difficult (Assmann, Pocock, Enos, & Kasten, 2000; Lagakos, 2006; Petticrew et al., 2012; Song & Bachmann, 2016; Wang, Lagakos, Ware, Hunter, & Drazen, 2007), but it is a step towards a personalised intervention effect, which can be computed using predictive modelling and has the potential to solve some of the challenges associated with an average treatment effect and subgroup analysis.

The process of randomly splitting data into training and test sets can transform the technical procedure of generative modelling into that of its predictive counterpart. In other words, we can also use data to train conventional regression models and then employ the trained models to predict outcomes on the test set. However, the assumptions made by conventional linear regression and new machine learning techniques we are going to encounter shortly are very different. The former assumes a linear and mathematical relationship between the outcome and observed features of the studied, whereas the latter takes data as the only input and allows the data to tell what that relationship really is (Popkin, 2015). Making connections between the "old" and "new" thus avoids defying the inferential contributions generative modelling has made.

To find out when some models shine and others struggle, we applied conventional logistic regression and some machine learning techniques to real and simulated datasets. We used logistic regression because the outcome is binary and it is easier to understand the percentage of accurate predictions than the mean squared errors when the outcome is continuous in an ordinary least squares regression. Nevertheless, the logic is the same, be it a classification or regression problem. The machine learning techniques