COMMENTARY

What failure to predict life outcomes can teach us

Filiz Garip[a,1]

Social scientists are increasingly turning to supervised machine learning (SML), a set of methods optimized for using inputs from data to forecast an unobserved outcome, to offer predictions to aid policy (1). Recent work scrutinizes this approach for its suitability to social science questions (2, 3) as well as its potential for perpetuating social inequalities (4). In PNAS, Salganik et al. (5) take a step back and ask a more fundamental question: are individual behaviors and outcomes even predictable?

Prediction in the Social Sciences

Prediction is not a typical goal in the social sciences despite recent arguments that it should be (6). Social scientists focus on inference: that is, understanding how an outcome is related to some input. The researcher selects a few inputs, specifies a parametric (often linear) model to connect inputs to the outcome, and estimates the parameters from data. The result is a simple and interpretable model that performs well in the sample at hand. In SML, by contrast, the researcher includes many inputs, considers flexible (often nonparametric) models linking inputs to the outcome, and picks the model that best predicts the outcome in new data. The result is a complex model that might perform well out of sample but often offers little insight into the mechanism linking inputs to the outcome.

Recent work connects these two cultures in different ways (7). First, researchers identify prediction tasks within the classical statistical framework and use SML to improve inference (8, 9). Second, scholars use predictions as a starting point to develop new theory (10). As SML becomes more mainstream in the social sciences and more students are drawn to seeking training in it, a crucial question remains unanswered. Apart from its use in well-delineated tasks, how helpful is SML in predicting human behaviors and outcomes?

Evaluating Predictability of Life Outcomes

Salganik et al. (5) offer a leap forward in evaluating the limits and promise of SML in the social sciences. The authors use the common task framework (11), where research teams compete on the same task, and launch a mass collaboration called the Fragile Families Challenge that is one of the first of its kind in the social sciences (12). The common task, in this case, is to predict six life outcomes in the Fragile Families study, a rich data source tracking over 4,000 families since the year 2000. The 160 research teams have full access to the first five waves of the study, which contain background variables, and partial access to the sixth wave, which measures outcomes to be predicted, such as children's grade point average or households' eviction experience. The teams use these data to train predictive models with any (SML or other) technique of their choice and are judged by a common metric (mean squared error) evaluating the accuracy of their predictions in the unseen part of the sixth wave that is available only to the challenge organizers.

The results are striking. The submitted models, regardless of the method used, all perform rather poorly in predicting the outcomes in the held-out data. Indeed, even the best predictions are not remarkably better than those offered by a simple benchmark model that employs standard regression models with a few expert-chosen inputs. The results produced by 160 independent teams using myriad strategies are clearly not an artifact of any one method and suggest that SML tools offer little improvement over standard methods in social science data.

This exercise has many lessons to teach us. First, if research teams went on their own to predict life outcomes with SML, their results (or nonresults) would probably never see the light of day. The common task method, in this instance, provides a solution to the well-known file drawer problem, whereby only positive results appear in scientific publications. Social scientists can use this method and follow the example set by the Fragile Families Challenge to ask fundamental questions of interest and to produce collective answers that are not bound by publication bias.

Second, the predictive framework shows us the importance of out-of-sample testing in social sciences,

[a]Department of Sociology, Cornell University, Ithaca, NY 14850

Author contributions: F.G. wrote the paper.

The author declares no competing interest.

See companion article, "Measuring the predictability of life outcomes with a scientific mass collaboration," 10.1073/pnas.1915006117.

[1]Email: fgarip@cornell.edu.

First published April 1, 2020.

even when our goal is not prediction per se. When we estimate and evaluate a model on the same sample, we run the risk of overfitting: that is, capturing the noise as well as the signal in data. The Fragile Families Challenge illustrates how humbling it can be to evaluate our models in new data. Whether our goal is prediction or not, we can use out-of-sample testing (or cross-validation procedures to efficiently partition data) to minimize overfitting (or p hacking) (13) and to assess the true performance of our models as well as the underlying theories (6). The recent push toward independent replication in social sciences represents a move in a similar direction (14, 15).

Third, the results from the challenge suggest what might be worth exploring further. There is a curious pattern in the data that requires more scrutiny. The prediction error from the competing models is strongly correlated with the family being predicted and only weakly related to the method being used. To put it differently, the predictions are accurate for most people in the data, regardless of the method used. However, the predictions are consistently off for a subset of families across all methods. This pattern begs the question: are we missing things essential to some families in our surveys? Is the problem, in other words, with our data as much as our methods?

Fourth, and the final point is also related to the data. The Fragile Families survey is a good representation of the kind of data typically available to social scientists in terms of its size and breadth. However, the data might be too limited for SML to truly shine. Future work needs to continue to evaluate these tools for the social sciences using larger datasets.

The Meaning of the Failure to Predict

To experts of SML, the unpredictability of life outcomes might come as a surprise, one that, for now, can be attributed to limitations of the Fragile Families data. However, what else could this observation mean? Salganik et al. (5) offer several useful insights. Our failure to predict life outcomes might mean that our survey measures (driven by our current theories) fall short of capturing relevant dimensions of people's lives, or it might mean that life outcomes are too idiosyncratic and subject to a predictability ceiling.

Salganik et al. offer a leap forward in evaluating the limits and promise of SML in the social sciences.

The fact that we cannot predict life outcomes, however, does not mean that we have little understanding of them. Our data and models might not allow us to forecast outcomes for each and every individual, but they can still help us produce aggregate descriptions [for example, of racial differences in school performance (16)] or identify causal relationships [such as the effect of education on earnings (17)]. Indeed, even if we were able to predict life outcomes with high accuracy, we would still want to isolate the mechanisms linking inputs to these outcomes, to ask counterfactual questions, and to design potential interventions. The predictive framework of the Fragile Families Challenge, in other words, does not replace or invalidate our standard inferential approach but rather, complements it.

Overall, the Fragile Families Challenge is a breakthrough in the social sciences for setting an example for mass collaboration and for evaluating the predictability of life outcomes in a high-quality longitudinal survey. The failure to predict individual and family outcomes, in this case, is anything but disappointing. It teaches us the value of collectively attacking a core question and subjecting our models to rigorous out-of-sample testing. It also reveals what our data (and the ideas on which they are based) might be missing and thus, charts a fruitful direction for future work.

1 J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, Prediction policy problems. *Am. Econ. Rev.* **105**, 491–495 (2015).
2 S. Mullainathan, J. Spiess, Machine learning: An applied econometric approach. *J. Econ. Perspect.* **31**, 87–106 (2017).
3 M. Molina, F. Garip, Machine learning for sociology. *Annu. Rev. Sociol.* **45**, 27–45 (2019).
4 S. Barocas, A. D. Selbst, Big data's disparate impact. *Calif. Law Rev.* **104**, 671–732 (2016).
5 M. J. Salganik *et al.*, Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 8398–8403 (2020).
6 D. J. Watts, Common sense and sociological explanations. *Am. J. Sociol.* **120**, 313–351 (2014).
7 L. Breiman, Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16**, 199–231 (2001).
8 S. Athey, G. Imbens, Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7353–7360 (2016).
9 J. Grimmer, S. Messing, S. J. Westwood, Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Polit. Anal.* **25**, 413–434 (2017).
10 N. Beck, G. King, L. Zeng, Improving quantitative studies of international conflict: A conjecture. *Am. Polit. Sci. Rev.* **94**, 21–35 (2000).
11 D. Donoho, 50 years of data science. *J. Comput. Graph. Stat.* **26**, 745–766 (2017).
12 E. L. Glaeser, A. Hillis, S. D. Kominers, M. Luca, Crowdsourcing city government: Using tournaments to improve inspection accuracy. *Am. Econ. Rev.* **106**, 114–118 (2016).
13 T. Yarkoni, J. Westfall, Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017).
14 J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
15 J. Freese, Replication standards for quantitative social science. *Sociol. Methods Res.* **36**, 153–172 (2007).
16 D. Downey, Black/white differences in school performance: The oppositional culture explanation. *Annu. Rev. Sociol.* **34**, 107–126 (2008).
17 D. Card, "The causal effect of education on earnings" in Handbook of Labor Economics, O. Ashenfelter, D. Card, Eds. (Elsevier, Amsterdam, the Netherlands, 1999), vol. 3A, pp. 1801–1863.