Productivity and Selection of Human Capital with Machine Learning

By AARON CHALFIN, OREN DANIELI, ANDREW HILLIS, ZUBIN JELVEH, MICHAEL LUCA, JENS LUDWIG, AND SENDHIL MULLAINATHAN*

* Chalfin: University of Chicago, Chicago IL 60637 (email: achalfin@uchicago.edu); Danieli: Harvard University, Cambridge MA 02138 (email: odanieli@fas.harvard.edu); Hillis: Harvard University, Cambridge MA 02138 (email: ahillis@fas.harvard.edu); Jelveh: New York University, New York NY 10003 (zjelveh@uchicago.edu); Luca: Harvard Business School, Boston, MA 02163 (email: mluca@hbs.edu); Ludwig: University of Chicago, Chicago IL 60637 and NBER (email: jludwig@uchicago.edu); Mullainathan: Harvard University, Cambridge MA 02138 and NBER (email: mullain@fas.harvard.edu). Paper prepared for the American Economic Association session "Predictive Cities," January 5, 2016. This paper uses data obtained from the University of Michigan's Inter-university Consortium on Political and Social Research. Thanks to the National Science Foundation Graduate Research Fellowship Program (Grant No. DGE1144152), and to the Laura and John Arnold Foundation, the John D. and Catherine T. MacArthur Foundation, the Robert R. McCormick Foundation, the Pritzker Foundation, and to Susan and Tom Dunn and Ira Handler for support of the University of Chicago Crime Lab. Thanks to Michael Riley and Val Gilbert for outstanding assistance with data analysis and to Edward Glaeser and Susan Athey for helpful suggestions. Any errors are our own.

Economists have become increasingly interested in studying the nature of production functions in social policy applications, Y = f(L, K), with the goal of improving productivity. For example what is the effect on student learning from hiring an additional teacher, $\partial Y/\partial L$, in theory (Lazear, 2001) or in practice (Krueger, 2003)? What is the effect of hiring one more police officer (Levitt, 1997)?

While in many contexts we can treat labor as a homogenous input, many social programs (and other applications) involve human services and so the specific worker can matter a great deal. Variability in worker productivity (e.g., Gordon, Kane and Staiger, 2006) means $\partial Y/\partial L$ depends on *which* new teacher or cop is hired. Heterogeneity in productivity also means that estimates for the effect of hiring one more worker are not stable across contexts – they depend on the institutions used to screen and hire the marginal worker.

With heterogeneity in labor inputs, economics can offer two contributions to the study of productivity in social policy. The first is standard causal inference around shifts in the level and mix of inputs. The second, which is the focus of our paper, is insight into selecting the most productive labor inputs – that is, workers. This requires *prediction*.

This is a canonical example of what we have called *prediction policy problems* (Kleinberg et al., 2015), which require different empirical tools from those common in micro-economics. Our normal tools are designed for causal inference – that is, to give us unbiased estimates for some $\hat{\beta}$. These tools do not yield the most accurate prediction, $\hat{Y}$, because prediction error is a function of variance as well as bias. In contrast, new tools from machine learning (ML) are designed for

prediction. They adaptively use the data to decide how to trade off bias and variance to maximize out-of-sample prediction accuracy.

In this paper we demonstrate the social-welfare gains that can result from using ML to improve predictions of worker productivity. We illustrate the value of this approach in two important applications – police hiring decisions and teacher tenure decisions.

I. Hiring Police

Our first application relates to efforts to reduce excessive use of force by police and improve police-community relations, a topic of great policy concern. We ask: By how much could we reduce police use of force or misbehavior by using ML rather than current hiring systems to identify high-risk officers and replace them with average-risk officers?

For this analysis we use data from the Philadelphia Police Department (PPD) on 1,949 officers hired by the department and enrolled in 17 academy classes from 1991-98 (Greene and Piquero, 2004). Our main dependent variables capture whether the officers were ever involved in a police shooting or accused of physical or verbal abuse (see Chalfin et al., 2016 for more details about the data, estimation and results). Candidate predictors come from the application data and include socio-demographic attributes (but not race/ethnicity), veteran or marital status, surveys that capture prior behavior and other topics (e.g., ever fired, ever arrested, ever had license suspended), and polygraph results.

We randomly divide the data into a training and test set, and use five-fold cross-validation within the training set to choose the optimal prediction function and amount by which we should penalize model complexity to reduce risk of over-fitting the data ("regularization.") The algorithm we select is stochastic gradient boosting, which combines the predictions of multiple decision trees that are built sequentially, with each iteration focusing on observations not well predicted by the sequence of trees up to that point (Hastie, Tibshirani and Friedman, 2008).

Figure I shows that de-selecting the predicted bottom decile of officers using ML and replacing them with officers from the middle segment of the ML-predicted distribution reduces shootings by 4.81% (95% confidence interval -8.82 to -0.20%). In contrast de-selecting and replacing bottom-decile officers using the rank-ordering of applicants from the PPD hiring system that was in place at the time would if anything *increase* shootings, by 1.52% (-4.81 to 5.05%). Results are qualitatively similar for physical and verbal abuse complaints.

One concern is that we may be confusing the contribution to these outcomes of the workers versus their job assignments – what we call *task confounding*. The PPD may, for example, send their highest-rated officers to the most challenging assignments, which would lead us to understate the performance of PPD's ranking system relative to ML. We exploit the fact that PPD assigns new officers to the same high-crime areas. Task-confounding predicts the ML vs. PPD advantage should get smaller for more recent cohorts – which does not seem to be the case.

II. Promoting Teachers

The decision we seek to inform is different in our teacher application: We try to help districts decide which teachers to retain (tenure) after a probationary period, rather than the decision we study in our policing application regarding whom to hire initially. Like previous studies, we find that a very limited signal can be extracted at hiring about who will be an effective teacher. Once people have been in the classroom, in contrast, it is possible to use a (noisy) signal to predict whether they will be effective.

Our data come from the Measures of Effective Teaching (MET) project (Bill and Melinda Gates Foundation 2013). We use data on 4th through 8th grade teachers in math (N=664) and English and Language Arts (ELA; N=707). We assume schools make tenure decisions to promote student learning as measured by test scores. Our dependent variable is a measure of teacher quality in the last year of the MET data (2011), Teacher Value-Add (TVA). Kane et al. (2013) leverage the random assignment of teachers to students in the second year of the MET study to overcome the problem of task confounding and validate their TVA measure.

We seek to predict future productivity using ML techniques (in this case, regression with a Lasso penalty for model complexity) to separate signal from noise in observed prior performance. We examine a fairly "wide" set of candidate predictors from 2009 and 2010, including measures of teachers (socio-demographics, surveys, classroom observations), students (test scores, socio-demographics, surveys), and principals (surveys about the school and teachers).

Figure II shows that the gain in learning averaged across all students in these school systems from using ML to deselect the predicted bottom 10% of teachers and replace them with average quality teachers was .0167$\boldsymbol{\sigma}$ for Math and .0111$\boldsymbol{\sigma}$ for ELA. The gains from carrying out this de-selection exercise using ML to predict future productivity rather than our proxy for the

current system of ranking teachers – principal ratings of teachers – equal .0072**σ** (.0002 to .0138) for Math and .0057**σ** (.0017 to .0119) for ELA.[1] Recall that these gains are reported as an average over all students; the gains for those students who have their teachers replaced are 10 times as large.

How large are these effects? One possible benchmark is the relative benefits and costs associated with class size reduction, from the Tennessee STAR experiment (Krueger, 2003). We assume following Rothstein (2015) that the decision we study here – replacing the bottom 10% of teachers with average teachers – would require a 6% increase in teacher wages. Our back-of-the-envelope calculations suggest that using ML rather than the current system to promote teachers may be on the order of two or three times as cost-effective as reducing class size by one-third during the first few years of elementary school.

III. General Lessons

In settings where workers vary in their productivity, using ML rather than current systems to hire or promote workers can potentially improve social welfare. These gains can be large both absolutely and relative to those from interventions studied by standard causal analyses in micro-economics.

Our analysis also highlights several more general lessons. One is that for ML predictions to be useful for policy they need to be developed to inform a specific, concrete decision. Part of the reason is that the decision necessarily shapes and constrains the prediction. For example, for purposes of hiring new police, we need a tool that avoids using data about post-hire performance. In contrast, for purposes of informing teacher tenure decisions, using data on post-hire performance is critical.

Another reason why it is so important to focus on a clear decision for any prediction exercise is to avoid what we call *omitted payoff bias*. Suppose an organization ranks workers using multiple criteria, but an algorithm predicts their performance on just one dimension. The use of that algorithm could potentially lead to a net *reduction* in the organization's goals (see Luca, Kleinberg and Mullainathan, 2016 for managerial examples). In general this will be less of an issue in situations where there is a strong positive correlation among all of the performance measures the organization cares about. Omitted payoff bias from predicting just a

[1] ML also shows gains relative to the TVA method by Kane et al. (2013), which controls for one-year lagged test scores and other factors. The ML versus TVA gains are 30-40% as large as the ML versus principal gains. We can also compare the ML approach to a model with two lags, which also yields an ML advantage – but these relative gains are smaller and, with the current data, not statistically significant. In some sense one contribution of ML in this case is to highlight the value of conditioning on the second lag of test scores.

single performance measure may be more of an issue if there are actually multiple dimensions of performance with some possible crowd-out among them, as in Holmstrom and Milgrom (1991).

In our applications omitted payoff bias will not be an issue for teacher promotion if one accepts achievement test scores as the key performance measure of policy concern, as many in the policy and research community seem to do, or believes that teaching other skills like critical thinking or creativity are complements to and not substitutes for teaching material covered by standardized tests. In our policing case the other dimension of productivity the public presumably cares about (beyond excessive use of force) is crime prevention, although our data do not include any direct measures of this. Whether crime prevention and risk of excessive use of force are positively or negatively correlated is not obvious. On the one hand, more proactive officers may initiate more citizen contacts, which may increase the risk of use of force. On the other hand police practices that enhance legitimacy in the eyes of residents may increase community cooperation with police to help solve crimes (Tyler and Fagan, 2008). These remain important open questions for future research to explore.

A final general lesson comes from the frequent challenge of having to train algorithms on data generated by past decisions, which can systematically censor the dependent variables (or "labels" in computer science). What Kleinberg et al. (2016) call the *selective labels problem* is most obvious in our police-hiring application, where we only have data on people the department actually hired. This means we cannot help select who to hire out of the original applicant pool, which could in principle let us reshuffle the ranking of current applicants in a way that could lead to gains in productivity at no cost. Instead with data on just those hired we can only inform a different decision – replacing predicted high-risk hires with average-risk officers – that would entail costs (higher wages to expand the applicant pool). Quasi-experimental variation in, say, how applicants are assigned to interview teams could help analysts determine whether department hiring decisions are based partly on information that is not made available to the algorithm.

There are many of these "picking people" applications for which ML prediction tools could be applied. Our goal with this paper is to stimulate more work on these problems.

REFERENCES

Bill and Melinda Gates Foundation (2013)

"Measures of Effective Teaching: 1 - Study Information." Inter-University Consortium for Political and Social Research, Ann Arbor, MI. ICPSR34771-v2. doi: http://doi.org/10.3886/ICPSR34771.v2.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig and Sendhil Mullainathan (2016) "Productivity and Selection of Human Capital with Machine Learniong." Harvard University Working Paper.

Gordon, Robert, Thomas J. Kane and Douglas O. Staiger (2006) "Identifying effective teachers using performance on the job." Hamilton Project Discussion Paper 2006-01.

Greene, Jack R. and Alex R. Piquero (2004) Supporting Police Integrity In Philadelphia Police Department, 1991-98 and 2000. Ann Arbor, MI: ICPSR.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2008) *The Elements of Statistical Learning, 2nd Edition*. Springer.

Holmstrom, Bengt and Paul Milgrom (1991) "Multitask principal-agent analyses: Incentive contracts, assert ownership and job design." *Journal of Law, Economics and Organization*. 7: 24-52.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller and Douglas O. Staiger (2013) "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." Seattle, WA: Bill and Melinda Gates Foundation.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig and Sendhil Mullainathan (2016) "Human decisions and machine predictions." Working Paper.

Krueger, Alan B. (2003) "Economic Considerations and Class Size." Economic Journal. 113(485): F34-63.

Lazear, Edward P. (2001) "Educational production." Quarterly Journal of Economics. 116(3): 777-803.

Levitt, Steven D. (1997) "Using electoral cycles in police hiring to estimate the effect of police on crime." American Economic Review. 87(3): 270-90.

Luca, Michael, Jon Kleinberg, and Sendhil Mullainathan (2016) "Algorithms need managers, too." *Harvard Business Review*. 104(1/2).

Rothstein, Jesse (2015) "Teacher Quality Policy When Supply Matters." *American Economic Review*. 105(1): 100-130.

Tyler, Tom R. and Jeffrey Fagan (2008) "Legitimacy and cooperation: Why do people help the police fight crime in their communities?" *Ohio State Journal of Criminal Law*. 6: 231-75.

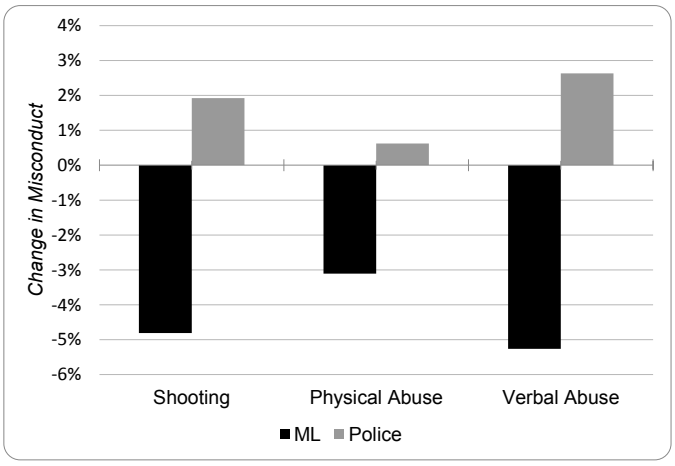

FIGURE 1. CHANGE IN POLICE MISCONDUCT

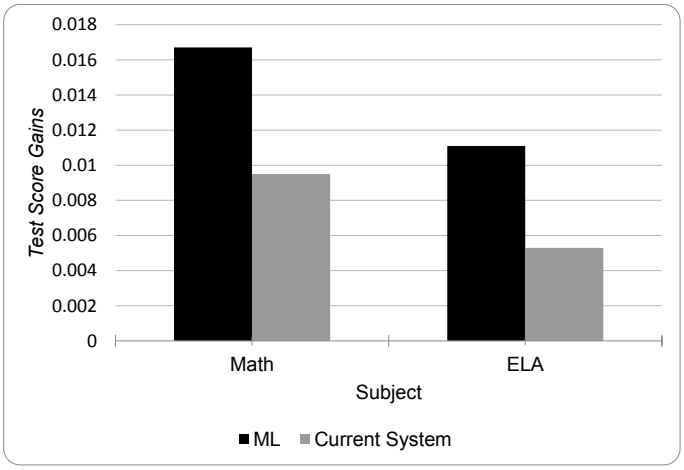

FIGURE 2. TEST SCORE GAINS, STANDARD DEVIATIONS