Supplemental Appendix for: Productivity and Selection of Human Capital with Machine Learning

Aaron Chalfin[1]
Oren Danieli[2]
Andrew Hillis[3]
Zubin Jelveh[4]
Michael Luca[5]
Jens Ludwig[6]
Sendhil Mullainathan[7]

January 11, 2016

[1]University of Chicago, achalfin@uchicago.edu
[2]Harvard University, odanieli@fas.harvard.edu
[3]Harvard University, ahillis@fas.harvard.edu
[4]New York University and University of Chicago, zjelveh@uchicago.edu
[5]Harvard Business School ,mluca@hbs.edu
[6]University of Chicago and NBER, jludwig@uchicago.edu
[7]Harvard University and NBER, mullain@fas.harvard.edu

Appendix A - Police Hiring

1 Introduction

In this appendix we provide additional details about our data and estimation procedures for predicting future officer performance with the Philadelphia Police Department (PPD) dataset (Greene and Piquero, 2006).

We are seeking to identify, using information available at the time a police officer is hired, which officers are most likely to engage in serious misconduct on the job in the future. In this application our proxy for serious misconduct refers to one of the following:

1. An incident leading to the provision of departmental discipline
2. An internal investigation
3. An off duty incident
4. An officer-involved shooting
5. Other misconduct
6. A citizen complaint for lack of service
7. A citizen complaint for physical abuse
8. A citizen complaint for verbal abuse

One limitation of the dataset we have available for analysis is that for many of these measures (such as physical or verbal abuse, or lack of service) the dataset captures allegations rather founded complaints. Another limitation is that it does not distinguish between police shootings that are founded versus unfounded.

For the analysis, we use data collected from the PPD on all 1,949 officers hired from 17 police academy cohorts between 1991-1998. While the available predictors are measured at the time the officer is hired, outcome variables are measured at a single point in time in 2000 and capture whether an officer has ever been accused of the above types of misconduct, but do not capture the *number* of misconduct accusations. For an outcome like a police shooting which is relatively rare, the binary indicator is likely to be highly correlated with the count of incidents. However, for an outcome like verbal abuse which is more common, the available data cannot be used to identify the most troublesome officers among the class of accused officers. Given this limitation, the number of events we estimate that could have been abated with the use of a prediction model is a lower bound.

2 Descriptive Statistics

We begin with a description of the incidence of each type of misconduct among the 1,949 officers in our sample — the means for each of these binary variables are presented in Table 1. These descriptive statistics provide the base rate against which our predictive model can be compared. Across all 17 academy cohorts, misconduct allegations are relatively rare, though they are naturally more common among the older cohorts which have the longest observation period and are less common among the most recent cohorts for whom on-the-job behavior is observed only for 1 or 2 years.

2.1 Outcome Variables

Overall, the most common type of complaint is for physical abuse (17 percent), followed by verbal abuse (10 percent), and lack of service (8 percent). It is worth noting that these are incidents that are made known to the police department when a citizen complains. It is easy to imagine that verbal abuse might in fact be the most common type of misconduct but might be less likely to be reported by citizens to the department. If citizen reporting is systematically related to officer characteristics, then our model would understate the risk of the most serious officers relative to the least serious officers. Officer-involved shootings (5 percent) and, to a lesser extent, off-duty incidents (10 percent) may provide a less confounded outcome to study as these incidents are far less sensitive to citizen reporting.

2.2 Predictors

The decision we are interested in informing is whether the PPD should have hired a given officer in the first place. While PPD has broad discretion over which officers to hire, once an officer is hired, and particularly once an officer completes a probationary period, it is comparatively difficult to terminate an officer. Hence we focus on factors that are known to PPD at the time of the hiring decision. The data are fairly rich and capture basic demographics (though we exclude race from the prediction model), prior military experience and work history including involuntary separation, prior drug use and criminal involvement and outcomes of polygraph testing. Table 2 shows summary statistics for the predictors used in our machine learning model.

Officers who are eventually hired are, on average, 27 years old at the time they apply to work as a Philadelphia police officer; 55 percent are nonwhite and one third are female. Seventeen percent of officers have served in the military and 2 percent are current members of the U.S. military reserve. Nearly all officers are registered to vote in Philadelphia, which reflects the fact that the city has a residency requirement in place for new officers. Officers have, on average, 13 years of schooling and have held an average of 5 prior jobs. Sixty-eight percent of officers have been unemployed in the past and over one quarter have been dismissed or fired. The majority of officers have never been arrested (84 percent), been a defendant in a criminal case (94 percent), been convicted of a crime (96 percent) or placed on probation pr parole (97 percent). Nearly half of officers report prior drug use but 100 percent of the drug use that officers reported during the hiring stage involved use of marijuana. The failure rate for polygraph tests was 28 percent for the sample.

3 Prediction Procedure

The features of standard econometric methods that make them less than ideal for purposes of generating accurate out-of-sample predictions are overcome by new methods in computer science and, in particular, the field of machine learning. Methods such as regularized (penalized) regression seek to balance the objectives of minimizing prediction bias and variance by minimizing a loss function that includes both in-sample fit and a term that penalizes more complex models that tend to increase variance when predicting out of sample. Similarly methods like tree-based classifiers make it possible to fully exploit increasingly rich data sources and to better understand the complicated interactions and non-linear relationships between large numbers of predictive factors.

A key lesson from the field of machine learning is that combining diverse signals can greatly improve prediction accuracy. For example the method of “ensemble training" (Hastie et al., 2009) involves generating a prediction model that combines multiple prediction models, including models that may capture limited signal and so have relatively weak predictive power when taken on their own. Previous research shows that the strength of the ensemble prediction model is enhanced when there is relatively more diversity in the underlying prediction models that are drawn upon (Kuncheva and Whitaker, 2003; Sollich and Krogh, 1996). The same logic applies to prediction models that are able to draw on diverse sources of data, each of which may be limited in its own way but by combining different data sources multiple types of signal can be combined and brought to bear on the prediction problem.

Recent work has demonstrated the success of ensemble decision-tree methods such as random forest and the related stochastic gradient boosting (Caruana and Niculescu-Mizil, 2006). Decision trees are nonparametric prediction models which split the input space of predictors into a set of rectangles where the splitting decision at each non-leaf node in the tree is based on a criterion such as squared loss. Each non-leaf node in the tree represents and if-then decision. A prediction can be generated for a new data point by dropping it down the tree and following the if-then decision path to a leaf node. The predicted value is a function, typically the mean for regression and majority vote for classification, of the outcome values of the subset of data points in the training set which ended up in the leaf node. One downside of decision trees is that they can be unstable; that is, the structure of these trees can vary considerably depending on slight chance changes to the dataset used to build the tree. To address this problem random forest builds many trees, where each tree is trained on a different bootstrapped sample of the original data and splitting decisions in tree nodes are based on a randomly sampled subset of the potential predictors. The final prediction model is a weighted average of the individual trees. These techniques help reduce model variance and prevent over-fitting.

Similar to random forest, gradient boosting Mayr et al. (2014) also grows an ensemble of trees but a key difference is that trees are developed sequentially. Intuitively, boosting works by iteratively upweighting difficult to predict data points. At any iteration t, the focus of a boosting algorithm is shifted to the mispredicted data points in the previous $t-1$ iterations. In the case of gradient boosting with a squared error loss function, each subsequent tree is essentially built to fit the previous trees' residuals. The estimator for the full model at any particular iteration is a weighted sum of all previous trees. The results reported here are from a set of gradient boosted tree models built for each misconduct outcome.

Given a response y and a set of predictors x_1 to x_k, the number of trees T, and a desired loss function (e.g. squared error for regression or Bernoulli deviance for classification), the gradient boosting algorithm works as follows:

Set initial guess to a constant value

For $t = 1, ..., T$

1. Evaluate the gradient of the loss function at each point to construct a new response
2. Fit a regression tree with K terminal nodes to the new response
3. Compute new predictions for each terminal node
4. Add the new tree to the ensemble at the learning rate γ

The number of terminal nodes K determines the number of interactions ($K-1$) allowed. Over-fitting can be avoided by reducing the influence of each new tree on the output of the model via a small value for γ in step 4, building short trees, randomly sampling a subset of the data in each iteration at step 2 to reduce prediction variance, and limiting the number of trees built. We select the optimal values for T, γ, and K via five-fold cross-validation.

4 Results

In this section we provide additional detail describing the results of our prediction models as well as statistical tests used to evaluate the quality of these models. We evaluate model quality using two benchmarking procedures: 1) a test of precision in the top decile of our prediction model relative to the departmental ranking established by PPD and 2) a counterfactual exercise in which we replace the officers who we identify as being *a priori* at greatest risk to commit misconduct with officers in the middle of the distribution.

In making these comparisons there is an implicit assumption that the PPD's ranking procedure was established solely to minimize misconduct. This may not be the case in practice, which as we note in our main paper text could lead to the problem of *omitted payoff bias*. However as we argue in the paper, recent public debates about officer misconduct make clear that there is no additional amount of police-officer productivity that the public is willing to tolerate in exchange for overlooking officer misconduct. In some sense the public's preferences over

police behavior seems to be lexicographic. Unfortunately the dataset we have available to us does not include any direct measures of "positive" officer productivity, so we cannot compare whether selecting officers based on ML predictions of misconduct yields a pool of officers with lower values of other productivity measures compared to PPD's current hiring system. This type of exercise, which would require different (and more) data, would be a valuable topic for future research.

4.1 Prediction Accuracy

Following standard practice in the machine learning literature, we report precision for each model — the proportion of officers in a given subsample of the data who actually commit misconduct. The advantage of focusing our attention on precision, one of several available metrics for assessing model quality, is that it is explicitly related to the decision a policymaker needs to make – which potential officers should the police department hire? We compute precision for each outcome by taking a weighted average of the class-level precision where the weights are equal to proportion of all officers in the top decile of a particular class. Table 4 shows that while neither ranking is highly predictive of misconduct, the machine learning ranking outperforms the PPD list across most outcomes and, as such, represents an improvement relative to the status quo.[1]

4.2 Precision

In order to establish that our prediction model is superior in identifying misconduct to the departmental ranking established by PPD, we motivate a formal test that compares precision between the two models. We focus here on a comparison of the top decile of predictions generated by our ML model and PPD's internal ranking.[2] Complicating this test is the fact that officers are independently ranked within 17 different academy classes and so, using the PPD ranking, we can only compare officers within but not between academy classes. We could test the null hypothesis that precision for the two models is equal for each academy class-outcome cluster but that would result in a series of severely underpowered tests as a given decile for an academy class-outcome cluster will consist of between 10-20 officers. Instead, we generate a stacked dataset in which, for a given outcome, we jointly test precision for all officers, conditional on academy class fixed effects that allow the base rate to vary by class. The test is outlined in model (1) below:

$$Y_{ij} = \alpha + \beta D_{ij} + \phi_j + \varepsilon_{ij} \tag{1}$$

In (1), Y_{ij} is a binary indicator variable indicating whether officer i in academy class j committed a particular type of misconduct. Each observation comes from either the top risk decile of either predictions from our ML model or in the PPD ranking. D_{ij} is a dummy variable indicating whether the observation is found in the ML ranking or the PPD ranking and ϕ_j are academy class fixed effects which allow each class to have a different base rate. β tests the null hypothesis that $\bar{Y}$ is equal between the ML and PPD samples. Because some officers (typically around 6 percent) will be in the top decile in both our ML model and the PPD ranking, we cluster standard errors on the officer to account for the fact that these are not completely independent samples.[3]

Results for this exercise can be found in Table 4. In the table, we provide precision at the top decile for both the ML model and the PPD ranking. Asterisks on the difference correspond to the significance of the result. Because the sample size across academy classes is fairly small (N 195), only two of the differences are significant at conventional levels — physical abuse ($p < 0.05$) and verbal abuse ($p < 0.10$). However, results for

[1] We also report the area under the curve (AUC) (Table 3) for the machine learning and PPD-generated risk rankings. The AUC can be interpreted as the probability that a randomly chosen officer with a misconduct will be predicted to be riskier than a randomly chosen officer without a misconduct. An AUC of 0.5 indicates that the ranking does no better than chance in discriminating between the two cases.

[2] The difference in precision between the two models is, in fact, greatest at the top decile compared to the nine other deciles. We argue that it is sensible to focus on the top 10 percent of the distribution which is where an outsize amount of misconduct occurs.

[3] In practice, due to the relatively small amount of overlap between the two samples, clustered and OLS standard errors are approximately identical.

all eight outcomes indicate that precision is higher for the ML model than the PPD ranking and p-values for several outcomes (internal investigations, off duty incidents and shootings) are below 0.15.

In order to provide an omnibus test for the equality of precision between the two models that is sufficiently well-powered, we pool across all outcomes in a single model that stacks outcomes for each officer in each academy class-outcome cluster, conditioning on interacted academy class $\times$ outcome fixed effects — thus allowing the base rate to differ by both academy class and outcome. This test is a natural extension of the test motivated in (1) except here we add an additional index, k, for the different outcome variables.

$$Y_{ijk} = \alpha + \theta D_{ijk} + \rho_{jk} + \epsilon_{ijk} \tag{2}$$

In (2), ρ_{jk} are now academy class $\times$ outcome fixed effects and θ tests whether $\bar{Y}$ differs across all outcomes. Results for this model are presented in the row entitled "stacked" in Appendix Table XX. Here, the raw difference in precision is 4.5 percent indicating that, within the top decile of the data, averaging across the 8 outcomes, the ML model correctly identifies nearly 5 percent more officers who have committed some type of misconduct than the PPD ranking. The result is significant at $p < 0.001$.

4.3 Changes in Misconduct

After establishing that our algorithm outperforms the PPD list at the top of the risk distribution in identifying a greater number of misconduct events, we next test whether use of the machine learning model could have reasonably led to a reduction in misconduct. To do so, we perform the following deselection exercise. We replace the top 10 percent of riskiest officers with a sample of officers drawn from the middle of the risk distribution in each academy class for both the PPD and machine learning list. We then compare the percentage change in misconducts between the new pool of officers versus the status quo. Figures 1, 2, and 4 show these results when the riskiest officers are replaced by those in the middle 33 percent, 50 percent, and 66 percent of the risk distribution, respectively. For nearly all outcomes and risk ranges, replacing risky officers using the machine learning ranking results in a reduction in misconducts. Replacement using PPD rankings tends to lead to an *increase* in misconducts.

To test whether these results are significant we use a bootstrapping procedure where, for each risk range, we perform the deselection exercise 1,000 times, resampling with replacement from the middle X percent of the risk distribution. Table 5 shows that when replacing with the PPD-generated list there is no significant change in the number of misconducts when compared to the status quo. For the machine learning list, there is a significant reduction in departmental discipline, physical abuse, verbal abuse, shooting, and other outcomes. The strength of the significance is stronger for departmental discipline and verbal abuse than for physical abuse, shooting and other outcomes.

4.4 Task Confounding

One potential concern with our approach is that we may confuse each person's contribution to productivity (misbehavior) from the contribution of the setting or job assignment — we refer to this problem as task confounding. Imagine, for instance, that PPD assigns its highest-rated officers to the most challenging beats in the city. This would create a relationship between each officer's unobservables and the outcome and would lead us to understate the predictive accuracy of the PPD's ranking system relative to that of the algorithm.[4]

To test for possible task confounding we exploit the fact that PPD assigns all new officers during their first year on the job (their probationary period) to similarly high-crime areas. So for the latest academy cohorts in our sample (those who started in 1998) the majority of time spent before the end of the study period in 2000 will be spent in essentially the same job assignment. More generally, the later the academy cohort, the less opportunity there is for this sort of task confounding. As a result, under the null hypothesis of task-confounding, the algorithm's advantage over PPD's system should be smaller for more recent cohorts.

[4]Figuring out the optimal way to assign officers to beats is beyond the scope of this exercise but we note that this may well be a first order issue in minimizing police misconduct.

We perform two tests of task confounding of the following form:

$$Advantage_{co} = \beta_0 + \beta_1 class + \beta_2 outcome \tag{3}$$

In the first test, $Advantage_{co}$ is defined as $\triangle_{ML} - \triangle_{PPD}$ where $\triangle_j$ is the percent change in the number of officers committing misconduct o in academy class c when the top risk decile is replaced by officers in the middle of the risk distribution using risking ranking $j \in \{ML, PPD\}$, $class$ is a time trend ranging from 1 to 17 and $outcome$ is a set of indicators for misconduct type. For the second test, $Advantage_{co}$ is defined as $ML_{prec} - PPD_{prec}$ where j_{prec} is the precision in the top decile for risk ranking j. If $\triangle_{ML} < \triangle_{PPD}$ ($ML_{prec} > PPD_{prec}$), then the algorithm is outperforming the police's ranking and a positive (negative) coefficient on $class$ implies that this advantage is decreasing over time, consistent with task confounding.

We estimate the significance of β_1 for the first test via a bootstrap approach where we sample with replacement at the class and outcome level 1,000 times. For each replication we estimate (3) and store up the value of β_1 and use the empirical distribution to assess significance. The first two columns of Table 6 show that there is no evidence that the algorithm systematically outperform the police rankings for earlier than late cohorts when the deselection exercise sampling is performed with the middle 60 percent of the risk ranking. Results are equivalent for the other risk ranges. For the second test, we alter our bootstrap approach where we sample with replacement the entire dataset 1,000 times and estimate the precision for the machine learning and PPD lists at the class and outcome level. We estimate (3) and examine the distribution of β_1. The last two columns of Table 6 again reveal no evidence that our algorithm performs better for earlier than later cohorts.

Appendix B - Teacher Tenure Decisions

1 Introduction

Our goal is to identify which teachers will best serve students. The primary outcome of interest is student performance on test scores. The standard way in which economists and policymakers pursue this objective is to develop a "value-add" measure—a score that reflects the effect a teacher has on a student's test scores relative to the average effect of all teachers.

Assume student i's test score is a function of some observable student characteristics X_i and section characteristics, X_s (for example, the prior test scores of students in her class). That is, we have:

$$Y_i = \beta X_i + \Gamma X_s + \varepsilon_{is} \tag{4}$$

Here ε_{is} captures other influences on student test scores apart from student and class characteristics. In particular, we will assume it includes a variable τ_s, the causal effect of that student's teacher on her test scores (relative to the average effect teachers have on students of this type) and other sources of noise, ν_{is}: $\varepsilon_{is} = \tau_s + \nu_{is}$. Since a teacher has many students, we can estimate τ with TVA, the mean of the residuals from the equation above.[5]

[5]This assumes that τ_s is orthogonal to $X_i$andX_s, an assumption that has been shown to be wrong. However, as most of the variation comes from within sections, this only mildly affects the estimation of the model.

$$
\begin{aligned}
TVA &= \tfrac{1}{N} \sum_i Y_i - \hat{Y}_i && (5)\\
&= \tfrac{1}{N} \sum_i \beta X_i + \Gamma X_{is} + \varepsilon_{is} - \beta X_i - \Gamma X_s && (6)\\
&= \tfrac{1}{N} \sum_i \varepsilon_{is} && (7)\\
&= \tau + \frac{1}{N} \sum_i \nu_{is} && (8)
\end{aligned}
$$

We now have a measure of teacher effectiveness, TVA, which includes (1) a true measure of teacher value-add, τ and (2) some noise, $\frac{1}{N}\sum_i \nu_{is}$. Two challenges arise:

1. Bias - Can we show that $E[\nu|\tau] = 0$? That is, are TVA measures unbiased on average?
2. Noise - Can we minimize $E[(TVA - \tau)^2]$? That is, can we reduce the noise in the estimate?

Most research has focused on (1), but (2) is very important if we want to rely on TVA as a measure of an individual teacher's effectiveness - for performance pay or teacher selection, for example.

2 Prior Work

We focus here on progress made by Kane et al. (2013) using data from the Measures of Effective Teaching Project and Chetty et al. (2014) using administrative data from a large urban school district.

2.1 Investigating Bias

Kane et al. focus on answering (1) above. Their data also come from the MET project. The main part of the project lasted two years. Kane et al. develop measures of TVA from the first year of the project. They then validate these measures in the second year, for which teachers were randomly assigned to classrooms. They find that their measure is an unbiased predictor of teacher contributions to student test scores in the second year.

Chetty et al. (2014) use a quasi-experimental research design and IRS data to show that the bias in similar teacher-value add measures is very small. The authors leverage teacher movements between schools in their research design. They show that on average, test scores in the relevant grade level of the new school shift as predicted by the change in teacher value add caused by the teacher's move. Moreover, conditional on prior test scores, TVA is almost uncorrelated with previously unobserved variables from IRS data about students' parents.

2.2 Handling Noise

Both papers handle noise in a similar manner. They use "shrinkage" of the estimators to reduce noise. Specifically, they use a lag model and regress TVA in year t on TVA in prior years.

$$
TVA_t = \sum_k \beta_k TVA_{t-k} \tag{9}
$$

The intuition is that TVA as measured in any given year will include noise, but by obtaining fitted values from a regression including prior years, we will reduce the noise uncorrelated across years.[6] The regressions

[6] Kane et al. (2013) also add teacher observational scores, experience and degree status.

produce coefficients significantly smaller than one. The fitted values can then be used as predictors for future TVA.

2.3 Gaps

While these papers establish that TVA is unbiased, their estimates of TVA still contain significant noise. This is evident from the reliability statistics reported in Chetty et al. (2014) and in a companion paper to Kane et al., Mihaly et al. (2013). Their findings suggest that using just one year, about half of the variation in TVA still comes from noise. Our calculations confirm this as well. This implies that the chances of a teacher truly at the bottom p percentile of TVA obtaining a predicted TVA score above the median is approximately p. For example, a teacher actually in the bottom 25th percentile of TVA would have roughly a 25% chance of being ranked above the 50th percentile.

These shortcomings are vexing for policy makers interested in using TVA to make decisions about individual teachers. Below, we outline how our approach with machine learning helps solve the problem of noise and produce predictions that can deliver better results for students.

3 Data

Our data come from the Measures of Effective Teaching (MET) project. The project collected potential indicators of teacher quality over two years and implemented a randomized controlled trial in the second year (AY 2010-2011) to validate them.

The data include student test scores; surveys of students, teachers, and principals; expert ratings of videos of teachers in the classroom; and assessments of teacher curriculum knowledge. We have tried different levels of aggregation of the data, but they did not seem to influence the results.

Six large school districts participated in the project, covering 317 schools, and more than 2,500 teachers from the fourth through ninth grade. Our analysis uses the subset of 4th-8th grade teachers present in both years of the data who participated in the randomized-controlled trial. We develop quality measures at the teacher-subject level (i.e. Math or ELA), yielding a sample size of 865 Math teachers and 922 English and Language Arts (ELA) teachers. Note that while our data is very wide (i.e. it has a large number of predictors), it is not very long (i.e. it does not have a very large number of observations).

To predict teacher performance, we use the measure of teacher quality previously validated from the MET data - Teacher-Value Add (TVA). For details on the construction and validation of this measure, see Kane et al. (2013).

4 Prediction Procedure

Our goal is to develop robust predictors of teacher performance in the following year. ML techniques allow the data to determine the right tradeoff between bias and variance for maximum predictive accuracy out of sample, allowing us to reduce the noise in measures of performance in the current year. A key element of this process is regularization – a technique that penalizes algorithms for choosing complex functions that likely over-fit the data, while retaining the ability to find signal among a wide set of variables and potential functional forms.

To find the optimal predictor of TVA, we first let the data select the optimal regularization parameter for a standard set of algorithms using 20-fold cross-validation. The algorithms we use include Random Forests[7],

[7]Random Forests produces its predictions by averaging over the predictions of multiple decision trees. Each tree is produced by searching over a random subset of all possible variables. A decision tree in general is built by iteratively choosing predictors that most separate the data according to an information criterion, constructing "layers" of the tree until some model complexity or minimum information gain is achieved.

Boosted Trees[8], Lasso[9] and Ridge.[10] The regularized linear algorithms we used performed similarly well and better than algorithms using decision trees.

We focus here on our results from using Lasso as it typically performs well in data sets with similar numbers of variables and observations like ours. Lasso is a form of regularized linear regression that imposes an L1-norm penalty on the coefficients, shrinking most coefficients to zero as a result. The final prediction for each teacher is produced with Lasso trained on a subsample that excludes that particular teacher.

We compare the results from Lasso to results from principal ratings. Principals were asked to rank their teachers relative to other teachers in the same grade/subject on a scale of one to six. We normalize this score to have zero mean and standard deviation of one, so that differences in scores across teachers are not due to differences across principals in the propensity to use high or low numbers on the scale. Put another way, each teacher's rating is relative to other teachers graded by the same principal, but on a scale comparable across principals.

We rank teachers by their predicted performance in 2011 (the following year) according to the Lasso and principal ratings. We then simulate replacing teachers predicted to be in the bottom of the distribution with average teachers (i.e. teachers whose TVA will be 0). We calculate confidence intervals using boot-strap for the difference in performance between the two prediction methods.

5 Results

Figure 5 shows the test score gains from replacing all teachers below particular thresholds of predicted performance for the next year with average teachers in the current year. Figure 1A shows the gains for Math; Figure 1B shows the gains for ELA. We narrow our sample to include only teachers with principal ratings in the data ($N = 664$ for Math and 707 for ELA).

Both figures demonstrate that the predictions from Lasso outperform the principal ratings across all thresholds. To consider a particular threshold, we compare results from deselecting 10% of teachers. The Lasso model yields test score gains of $.0167\sigma$ for Math and $.0111\sigma$ for ELA[11]. The principal ratings yield gains of $.0095\sigma$ for Math and $.0053\sigma$ for ELA. That is, for deselecting 10% of teachers, the predictions produced by machine learning create an improvement of $.0072\sigma$ for Math and $.0057\sigma$ for ELA over the principal ratings.

We also compare the performance of the Lasso model to the estimates developed by Kane et al. (2013) from the MET data. The predictions from that model are produced using one lag of TVA and a few teacher characteristics. Deselecting 10% using that model yields a $.0138\sigma$ gain for Math and $.0094\sigma$ gain for ELA. The Lasso model thus adds an improvement of $.0029\sigma$ for Math and $.0017\sigma$ for ELA over the Kane et al. estimates (roughly 30-40% of the improvements of the Lasso model over the principal ratings). These improvements are nearly statistically significant at the 5% level.

Finally, we compare performance relative to predictions from a model that includes two lags of TVA, similar to estimates produced in Chetty et al. (2014). The Lasso model still improves performance by 0.0010σ for Math (95% confidence interval of -.0020 to .0051) and $.0004\sigma$ for ELA (-.0055 to .0056).

Overall, the machine learning approach outperforms other models in predicting TVA (albeit not always statistically significantly). It is worth noting that these improvements would be expected to increase with more data - for both longer and wider data.

[8] Boosted Trees weights multiple trees to form a prediction. In this algorithm, the trees are built sequentially and the data fed to each iteration is re-weighted according to the previous performance on that data.

[9] Lasso and Ridge are forms of "regularized" regression, in which the coefficients are chosen to minimize the sum of (1) the squared errors and (2) the size of the coefficients, weighted by a regularization parameter. In Lasso, (2) enters as the sum of the absolute values of the coefficients (the L1 norm). In Ridge, (2) enters as the sum of the squared values of the coefficients (the L2 Norm).

[10] For a more detailed explanation of machine learning and the algorithms discussed above, see Abu-Mostafa et al. Learning from Data (2012) and Bishop Pattern Recognition and Machine Learning (2007).

[11] Note that these are improvements for all students. For those students whose teacher is replaced, the gains are $.167\sigma$ for Math and $.111\sigma$ for ELA.

For traditional econometric techniques such as OLS, longer data (i.e. higher N) will merely reduce the variance of estimated parameters. In contrast, longer data allows machine learning to develop richer models, capturing more nuanced behavior of the data while minimizing the risk of overfitting. In part, this explains why the machine learning approach in this project yielded a linear model. A more complex model would likely need more data. The length of this data set is small relative to most ML projects. For these reasons, longer data would raise the improvement from machine learning over the traditional approach.

Wider data (i.e. more variables) also raises the improvement from machine learning. OLS has a tendency to over fit the data it receives; with more variables, it's possible that OLS will even perform worse by fitting noise within the sample. Regularization in the machine learning approach avoids fitting the noise, while allowing the algorithm to find signal in a wider set of variables. Given results from Dobbie (2011), who finds predictive power in personality traits measured by Teach For America, we believe that there is more signal to be found - and machine learning is well placed to sort it from noise.

6 Cost Benefit Analysis

We compare the cost effectiveness of deselecting teachers with machine learning to another intervention in K-12 education, decreasing class size.

The increase in costs associated with stricter tenure decisions comes from the need to compensate teachers for a system in which tenure is a riskier prospect. Rothstein (2015) uses a structural model to estimate this amount. His model shows that to replace the bottom 20% of teachers, we would need a 12% increase in wages to keep the same number of teachers. An approximate required increase in wages for deselecting the bottom 10% is therefore 6%. In AY 2011-2012, the instruction cost per student in US public schools was $6,706, leading to a required cost increase of $402 per student.[12]

To calculate the costs of a decrease in class size, we use the same calculation as in Krueger (2003) for the STAR experiment. The STAR experiment increased the number of classes by 47%. The average student spent 2.3 years in the experiment. The present value of costs to decrease class size in this example is given by:

$$PV = C_t + \frac{C_t}{1+r} + 0.3\frac{C_t}{(1+r)^2} \tag{10}$$

Based on a total expenditure per student of $11,014 in AY2011-12 we get that $C_t = \$5177$ and for $r = .04$, $PV = \$11,590$.

The STAR experiment yielded an improvement in test scores of .15 standard deviations, implying a cost to benefit ratio of .0129 standard deviations / $1K. In contrast, deselecting teachers yields a ratio of .0415 for Math and .0276 for ELA. That is, our approximate calculations imply that deselection using machine learning is 2-3 times as cost effective as decreasing class size.

[12] See `https://nces.ed.gov/programs/coe/indicator_cmb.asp`

References

Caruana, Rich and Alexandru Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in "Proceedings of the 23rd international conference on Machine learning" ACM 2006, pp. 161–168.

Chetty, Raj, John N Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 2014, *104* (9), 2593–2632.

Dobbie, Will, "Teacher Characteristics and Student Achievement: Evidence from Teach For America," *mimeo*, 2011, (July).

Greene, Jack R. and Alex R. Piquero, "Supporting Police Integrity in the Philadelphia [Pennsylvania] Police Department, 1991-1998 and 2000," 2006.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning* Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2009.

Kane, Thomas J, Daniel F Mccaffrey, Trey Miller, and Douglas O Staiger, "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment," *Report for the Measures of Effective Teacher Project*, 2013.

Krueger, Alan B., "Economic considerations and class size," *The Economic Journal*, 2003, *113* (485), F34–F63.

Kuncheva, Ludmila I. and Christopher J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine Learning*, May 2003, *51* (2), 181–207.

Mayr, A., H. Binder, O. Gefeller, and M. Schmid, "The Evolution of Boosting Algorithms: From Machine Learning to Statistical Modelling," *Methods of Information in Medicine*, August 2014, *53* (6), 419–427.

Mihaly, Kata, Daniel F Mccaffrey, Douglas O Staiger, and J R Lockwood, "A Composite Estimator of Effective Teaching," *MET Project*, 2013.

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

Sollich, Peter and Andres Krogh, "Learning with Ensembles: How over-fitting can be useful," in "Advances in Neural Information Processing Systems" MIT Press 1996, pp. 190–196.

Outcome Variables	Mean	SD
Physical Abuse (ever)	0.17	0.37
Verbal Abuse (ever)	0.1	0.3
Lack of Service (ever)	0.08	0.28
Internal Investigations (ever)	0.15	0.36
Off Duty Incidents (ever)	0.1	0.3
Police Shootings (ever)	0.05	0.22
“Other" Misconduct	0.08	0.28
Departmental Discipline	0.3	0.46

TABLE 1: Summary statistics for outcome variables

Predictors	Mean	SD
Age	26.71	6.19
Race (1 = non-white)	0.55	0.5
Sex (1=female)	0.33	0.47
Recipient of veteran's Preference (1=yes)	0.08	0.38
Military Reserve Status (1=yes)	0.02	0.13
Number of Files	1.16	0.49
Total # of addresses in past 10 years	3.42	2.28
Registered voter in Philadelphia (1=yes)	0.95	0.21
Registered voter in other location (1=yes)	0.03	0.16
Has valid PA driver's license (1=yes)	1	0.02
PA license ever suspended (1=yes)	0.2	0.4
License from other state suspended (1=yes)	0.02	0.14
Ever involved in car accident (1=yes)	0.67	0.47
Total # of car accidents involved in	1.51	0.86
Received traffic ticket in past 5 years (1=yes)	0.39	0.49
Total # of traffic tickets issused	1.54	1.07
Years of schooling	13.24	1.77
Total # of schools listed	4.6	1.83
Total # of prior jobs	5.23	2.71
Any length of unemployment? (1=yes)	0.68	0.47
Ever been dismissed / fired (1=yes)	0.28	0.45
Ever been dismissed from organization (1=yes)	0.02	0.13
Ever member of violent organization (1=yes)	0	0.04
Behind on bills? (1=yes)	0.28	0.45
Any loans, debt in excess of $1000 (1=yes)	0.66	0.47
Total amount owed-consumer debt (USD)	5972.19	8452.52
Ever filed for bankruptcy (1=yes)	0.03	0.16
Under court order? (1=yes)	0.07	0.26
Ever member of military (1=yes)	0.17	0.38
Type of military discharge	0.03	0.17
Any parent a police officer (1=yes)	0.11	0.31
Total # of children	0.94	1.28
Total # of siblings	3.2	2.3
Total # of family members arrested	0.6	0.97
Currently charged with any crime (1=yes)	0	0.05
Currently on probation/parole (1=yes)	0	0
Presently free on bail or ROR (1=yes)	0	0
Currently wanted on outstanding warrant (1=yes)	0	0.03
Currently subject of a protection from abuse complaint (1=yes)	0	0.05
Currently under indictment (1=yes)	0	0
Ever interviewed or questioned by law enforcement? (1=yes)	0.53	0.5
Ever been arrested? (1=yes)	0.16	0.36
Ever been convicted of a crime? (1=yes)	0.04	0.2
Ever on probation / parole? (1=yes)	0.03	0.16
Ever paid a fine? (1=yes)	0.37	0.48
Ever had to pay restitution? (1=yes)	0.02	0.15
Ever had to pay any court cost? (1=yes)	0.1	0.3
Ever had to post bail? (1=yes)	0.02	0.15
Ever lost or forfeited posted bail? (1=yes)	0	0.05
Ever been a defendant in a criminal case (1=yes)	0.06	0.23
Ever been questioned about a crime (1=yes)	0.25	0.43
Ever received a subpoena (1=yes)	0.26	0.44

Predictors	Mean	SD
Ever plead "no contest" to a criminal charge (1=yes)	0.02	0.13
Ever had police come to your house to investigate a crime (1=yes)	0.21	0.41
Ever been subject of a protection from abuse order (1=yes)	0.03	0.17
Ever been subject of a criminal complaint (1=yes)	0.02	0.15
Ever been a character witness (1=yes)	0.04	0.2
Ever been investigated for child abuse / neglect(1=yes)	0.03	0.16
Ever been investigated for spousal abuse (1=yes)	0	0.03
Ever applied for LE job? (1=yes)	0.51	0.5
Total # of times not hired	0.91	1.5
Ever been a member of PPD or other lae enforcement agency? (1=yes)	0.09	0.29
Ever applied for job with city of Philadelphia (1=yes)	0.45	0.5
Total # of times not hired for city job	0.53	0.85
Ever used solvents, inhalants, etc. (1=yes)	0.05	0.21
Ever sold solvents, inhalants, etc. (1=yes)	0.04	0.19
Ever sold or given prescription drugs (1=yes)	0.21	0.41
Possessed marijuana past 6 months (1=yes)	0.02	0.14
Ever possessed marijuana (1=yes)	0.45	0.5
Used marijuana past 6 months (1=yes)	0	0.03
Ever used marijuana (1=yes)	0.48	0.5
Ever possessed any illegal drug (1=yes)	0.46	0.5
Ever purchased any illegal drug (1=yes)	0.14	0.35
Ever "chipped-in" to purchase illegal drug (1=yes)	0.04	0.2
Ever used any illegal drug (1=yes)	0.48	0.5
Ever been present when someone else used an illegal drug (1=yes)	0.86	0.35
Ever sold any type of illegal drug (1=yes)	0.2	0.4
Now or ever owned firearm (1=yes)	0.25	0.43
Ever obtained a permit to carry firearm (1=yes)	0.12	0.33
Polygraph - total number of times taken	1.54	0.99
Polygraph - total # of times deception indicated	0.43	0.88
Polygraph - total # of inconclusive tests	0.02	0.15

TABLE 2: Summary statistics for predictor variables

	AUC	
	PPD	ML
Dept. Discipline	0.5560695	0.5701806
Internal Investigation	0.5552798	0.5440366
Lack of Service	0.5683092	0.5755264
Off Duty	0.5349831	0.5832576
Other	0.5899378	0.5882592
Physical Abuse	0.582911	0.5617264
Shooting	0.5881667	0.6350581
Verbal Abuse	0.5515384	0.6066207

TABLE 3: Area under the curve for the machine learning and PPD rankings.

	Precision PPD	Precision ML	Difference	p-value
Stacked	0.119	0.165	0.045***	0.001
Departmental Discipline	0.296	0.367	0.071	0.100
Internal Investigations	0.143	0.173	0.030	0.370
Lack of service complaints	0.092	0.097	0.005	0.848
Off duty incidents	0.071	0.117	0.046	0.117
Physical abuse complaints	0.138	0.224	0.086**	0.020
Verbal Abuse complaints	0.087	0.143	0.056*	0.054
Shootings	0.046	0.082	0.036	0.144
Other	0.082	0.112	0.030	0.303

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

TABLE 4: Outcome-level and stacked precision.

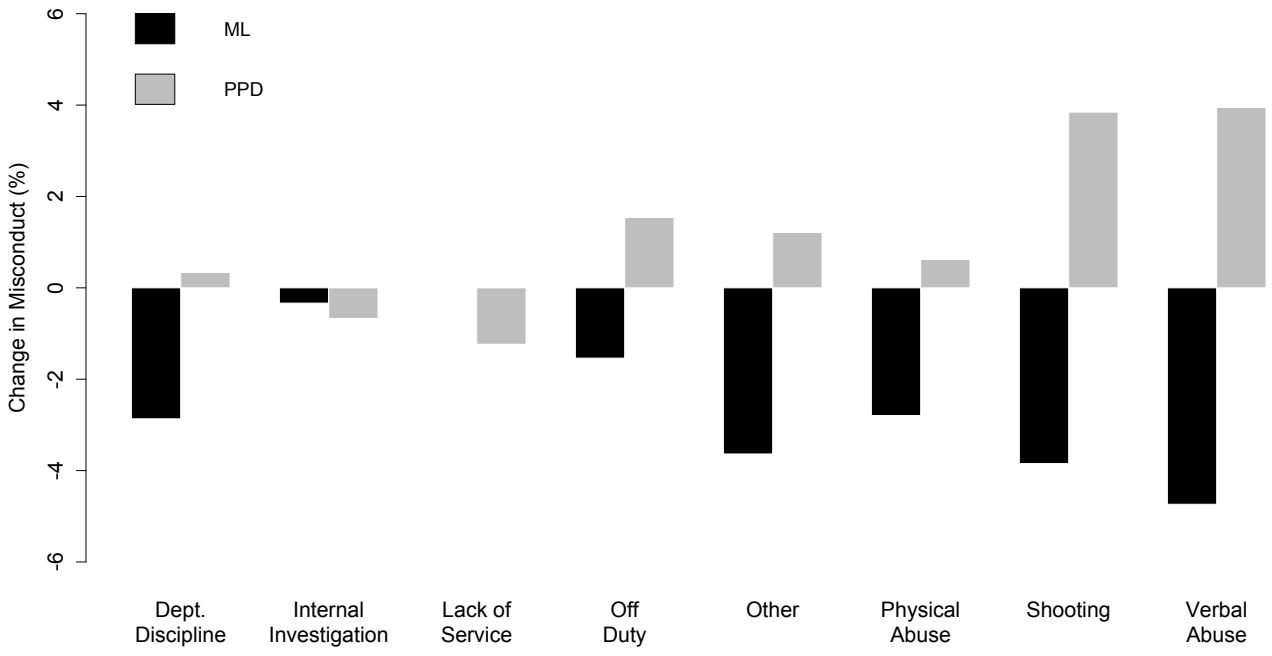


FIGURE 1: Change in misconducts when deselecting by sampling from the middle 33% of the risk disrtibution.

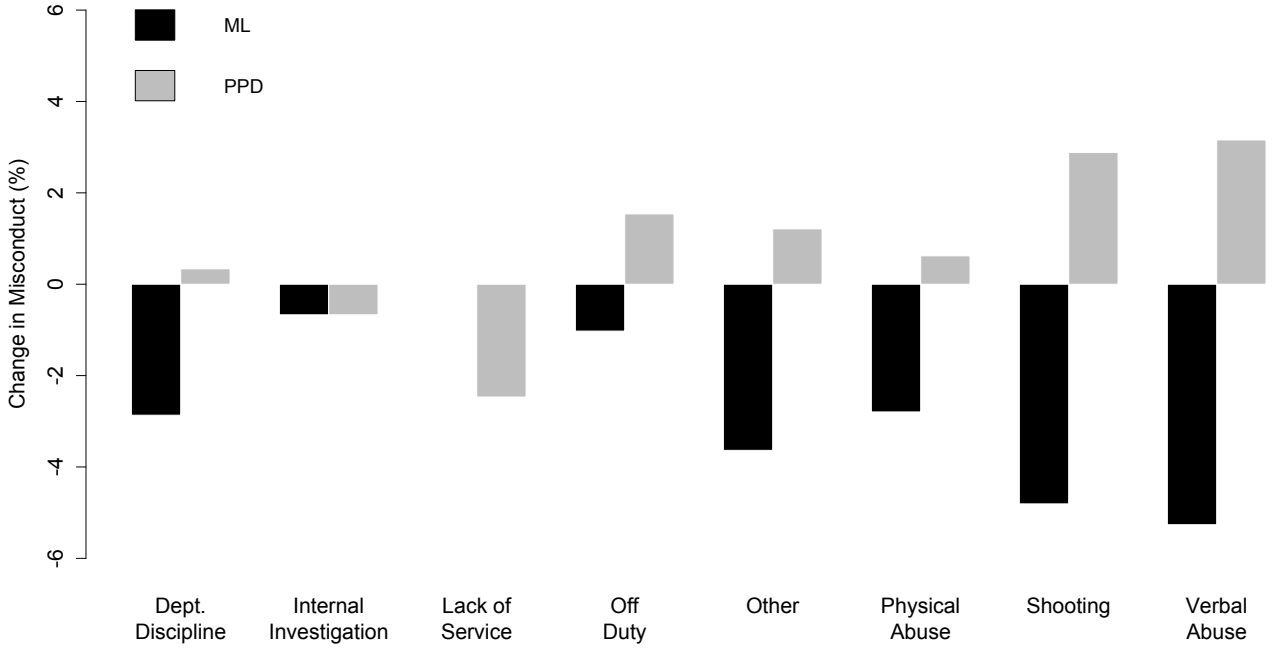


FIGURE 2: Change in misconducts when deselecting by sampling from the middle 50% of the risk disrtibution.

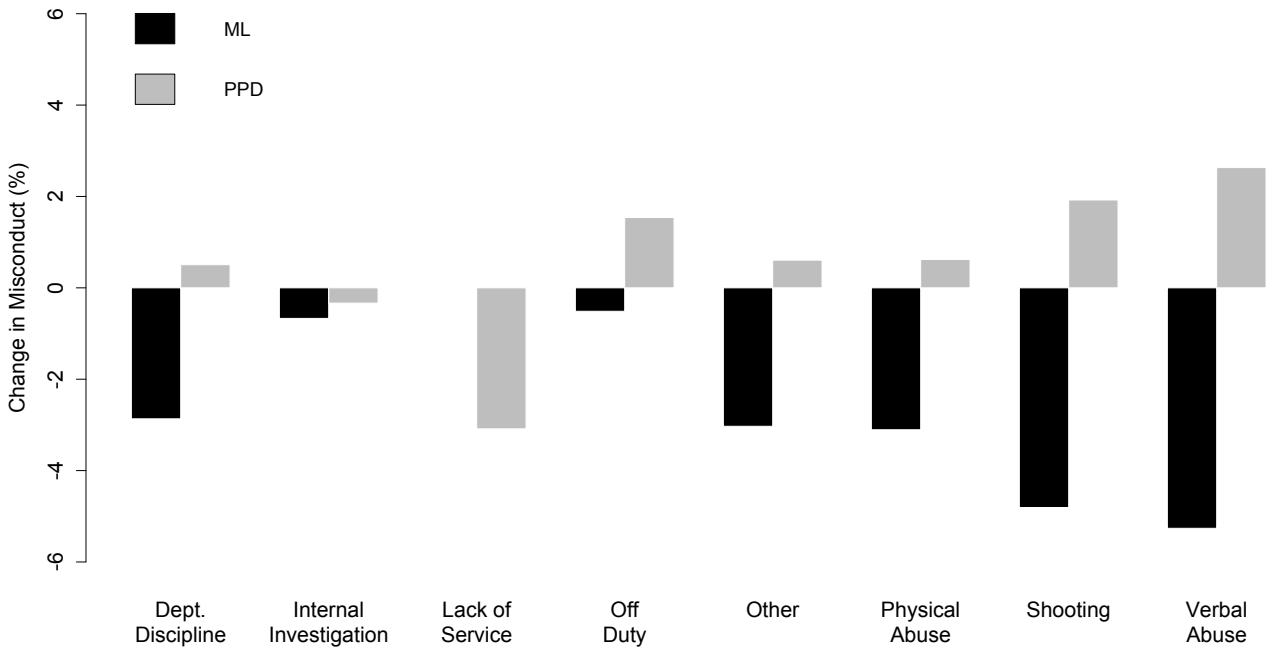


FIGURE 3: Change in misconducts when deselecting by sampling from the middle 66% of the risk disrtibution.

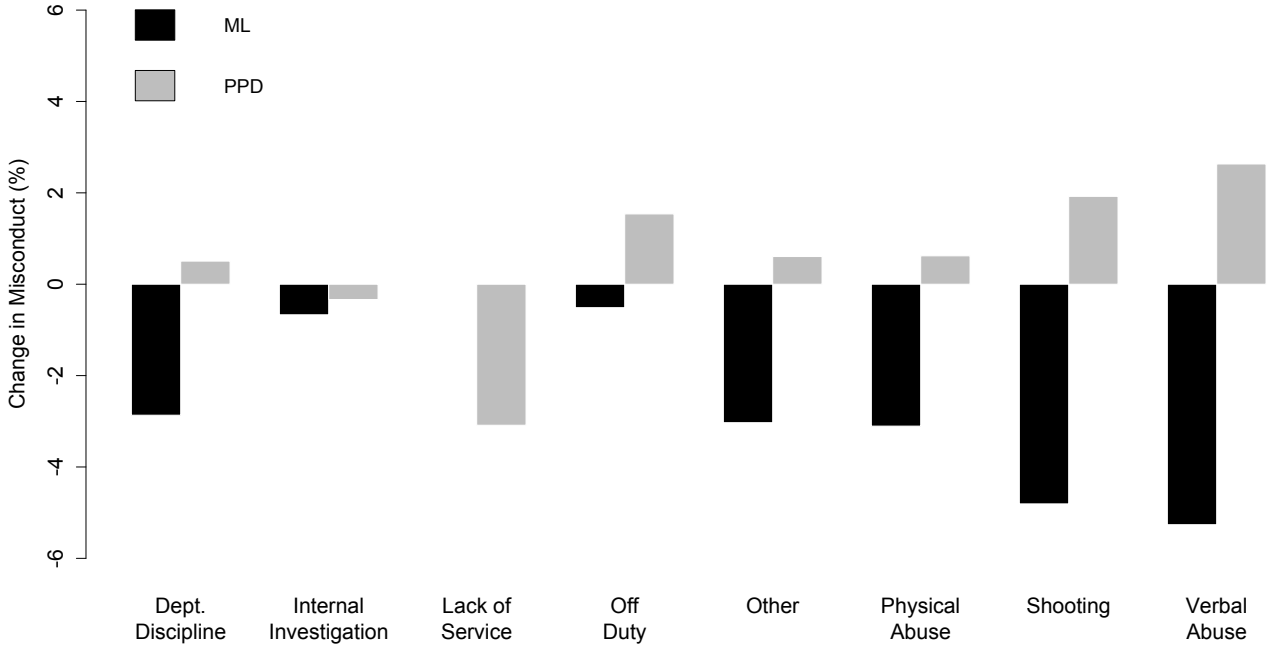


FIGURE 4: Change in misconducts when deselecting by sampling from the middle 66% of the risk disrtibution.

Comparison	ML vs. Status Quo			PPD vs. Status Quo			ML vs. PPD		
Risk Range	33%	50%	66%	33%	50%	66%	33%	50%	66%
Dept. Discipline	-0.029***	-0.029***	-0.029***	0.003	0.003	0.005	-0.032**	-0.030**	-0.034**
Internal Investigation	-0.003	-0.007	-0.007	-0.007	-0.007	-0.003	0.007	0.003	0.000
Lack of Service	0.000	0.000	0.000	-0.012	-0.025	-0.031	0.012	0.025	0.025
Off Duty	-0.015	-0.01	-0.005	0.015	0.015	0.015	-0.031	-0.023	-0.021
Other	-0.036*	-0.036*	-0.03	0.012	0.012	0.006	-0.048*	-0.048*	-0.036
Physical Abuse	-0.028	-0.028*	-0.031**	0.006	0.006	0.006	-0.034*	-0.034*	-0.040*
Shooting	-0.038*	-0.048*	-0.048*	0.038	0.029	0.019	-0.077*	-0.077**	-0.067**
Verbal Abuse	-0.047**	-0.053**	-0.053**	0.039	0.032	0.026	-0.089**	-0.079***	-0.079***

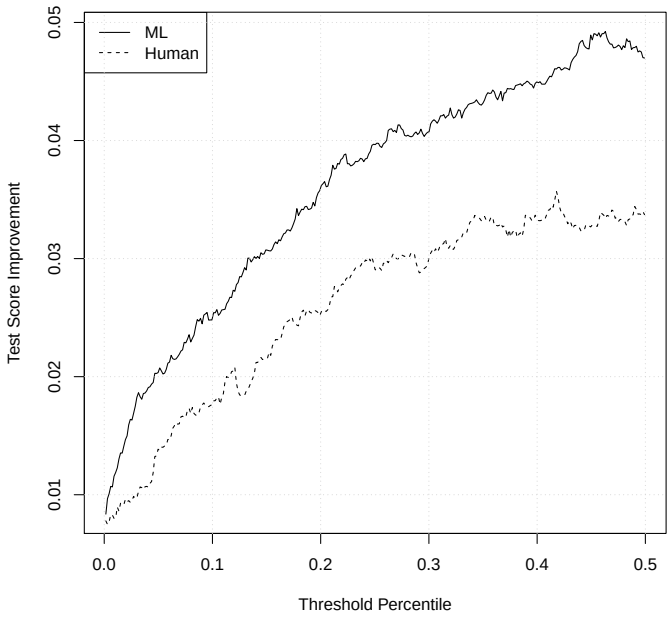
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

TABLE 5: The change in misconducts at the median from 1,000 bootstrap samples of the deselection exercise. The first set of results compare the change in misconducts between the machine learning-altered list and original list, the second set for PPD-altered list and the original list, and the third set for the machine learning-altered and the PPD-altered list.

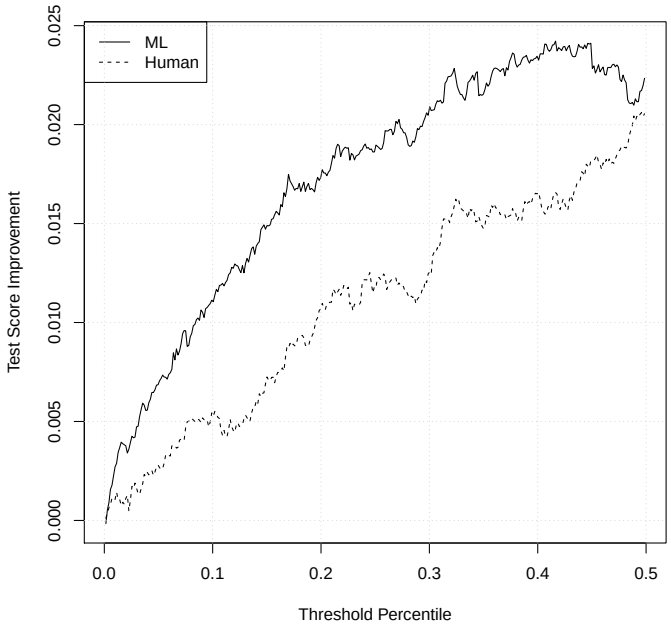
Outcome	$\triangle_{ML} - \triangle_{PPD}$	$\triangle_{ML} - \triangle_{PPD}$	$ML_{prec} - PPD_{prec}$	$ML_{prec} - PPD_{prec}$
$class$	-0.0052^{*}	0.0039	-0.0022	0.0065
	(0.0030)	(0.0111)	(0.0024)	(0.0104)
$class^2$		-0.0005		0.0065
		(0.006)		(0.0104)

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

TABLE 6: Tests for task confounding.



(A) Math



(B) ELA

FIGURE 5: Test Score Gains