Machines Learning Justice: A New Approach to the Problems of Inconsistency and Bias in Adjudication

Hannah Laqueur and Ryan Copus[*]

May 22, 2016

Abstract

We offer a two-step algorithmic approach to the problems of inconsistency and bias in legal decision making. First, we propose a new tool for reducing inconsistency: Judgmental Bootstrapping Models ("JBMs") built with machine learning methods. JBMs, by providing judges with recommendations generated from statistical models of themselves, can help those judges make better and more consistent decisions. To illustrate these advantages, we build a JBM of release decisions for the California Board of Parole Hearings. Second, we describe a means to address systematic biases that are embedded in an algorithm (e.g., disparate racial treatment). We argue for making direct changes to algorithmic output based on explicit estimates of bias. Most commentators concerned with embedded biases have focused on constructing algorithms without the use of bias-inducing variables. Given the complex ways that variables may correlate and interact, that approach is both practically difficult and harmful to predictive power. In contrast, our two-step approach can address bias without sacrificing performance.

[*] PhD Candidates, University of California, Berkeley Law School, Jurisprudence and Social Policy. The authors shared equally in authorship and are listed in reverse alphabetical order. We thank Jennifer Shaffer and Howard Moseley of the California Board of Parole Hearings for many patient conversations, openness to analysis, and assistance with data. We similarly thank Vanessa Nelson-Sloan and the Life Support Alliance for being a fruitful and reliable source of knowledge. We are also extremely grateful to for the guidance and comments from Christopher Elmendorf, David Hoffman, Rob MacCoun, Anne Joseph O'Connell, Kevin Quinn, Sarath Sanga, and Franklin Zimring. In addition, we would like to thank our research assistants Brian Ashworth, Kevin Chen, Chance Grable-Hughes, Karalyn Lacey, Shivani Pandare, and Wenjie Xu. Finally, we thank Jonathan Simon and the Berkeley Empirical Legal Studies Fellows for their comments and feedback.

Contents

Introduction 3

1 The Problems of Inconsistency and Bias 5

2 Addressing Inconsistency with Judgmental Bootstrapping 7

2.1 Existing Literature on Judgmental Bootstrapping 7

2.2 A New Conceptual Framework: Crowdsourcing and Condorcet's Jury Theorem 8

2.3 Machine-Learning JBMs 10

3 Illustrating Judgmental Bootstrapping with California Parole Decisions 11

3.1 Building and Describing the JBM 13

3.1.1 JBM Predictions 13

3.2 Using the JBM to Address Inconsistency 15

3.2.1 Implementation Details 16

3.2.2 Evidence of Better Release Decisions 18

4 Adjusting the JBM to Address Bias 19

5 Additional Considerations 22

5.1 Integrating JBMs with the Rise of Evidence-Based Adjudication 22

5.2 Some Pragmatic Concerns 25

6 Conclusion 28

A Appendix 35

A.1 Super Learner 35

Introduction

"Justice," the adage goes, "is what the judge ate for breakfast." The problems of inconsistency and bias in legal decision making have long been a concern. Recent research has documented substantial inconsistencies within and across judges. It suggests, for example, that the outcome of a football game (Chen and Spamann 2014), the results of the immediately preceding case (Chen, Moskowitz, and Shue 2014), and the time of day (Danziger, Levav, and Avnaim-Pesso 2011) can substantially affect legal decisions. Wide between-judge disparities have been found in domains including immigration asylum (Ramji-Nogales, Schoenholtz, and Schrag 2007; Fischman 2014), social security disability (Nakosteen and Zimmer 2014), and criminal sentencing (Abrams, Bertrand, and Mullainathan 2012). Although identifying and measuring bias, particularly racial bias, has proven more difficult to document, it is nonetheless a major concern and subject the subject of a great deal of research.

In this paper, we propose extending the research on judgmental bootstrapping (Dawes 1971) to develop a new two-step approach to the problems of inconsistency and bias in legal decision making. Specifically, we advocate the use of machine learning to build Judgmental Bootstrapping Models ("JBMs") that, by modeling decision-makers, can generate holistic recommendations that help judges mitigate inconsistency and make better decisions. Despite being built to copy human decisions, theory and research suggests these decision-predictive algorithms (JBMs) can actually improve upon human judgments by effectively crowdsourcing across and within decision makers, thereby canceling out arbitrary and contingent factors and minimizing inconsistency and random biases. In a second step, JBM output can be adjusted to address systematic biases such as disparate racial treatment. By focusing on the output rather than the variables included in a model, this two-step approach can address bias without sacrificing predictive performance.

In contrast to the existing literature on JBMs, which has used multivariate regression to model an expert's reasoning process, we advocate developing machine-learning JBMs that mimic human intelligence only in results. This machine-learning approach is distinct in method, aim, and consequence to the traditional decision-rule JBMs.[1] The aim of the machine-learning approach is not to model the expert's decision rules, but instead to produce recommendations that make it *seem* as if it has discovered the expert's decision rules. While the efficacy of both approaches relies on the ability to reduce noise and focus on the signal, in contexts in which decisions require nuanced distinctions, the decision-rule approach may remove too much signal along with the noise.

The paper proceeds in five main parts. Section 1 briefly discusses the problems of inconsistency and bias in legal decision making. Section 2 provides a conceptual account of judgmental bootstrapping and its benefits. We introduce readers to the judgmental bootstrapping literature and give a new theoretical account of JBMs' ability to mitigate inconsistency in a way that can lead to better decisions. We argue for a machine learning, decision-based approach to building JBMs as compared to the traditional multivariate regression, process-based approach. Section 3 offers an analysis of the California Board of Parole suitability hearings to illustrate the potential of JBMs to mitigate inconsistency and improve decision-making. We use a newly constructed dataset to build and analyze a predictive model of the California Board of Parole Hearings' release decisions. We use our JBM of release decisions, built using the machine-learning ensemble approach "Super Learner" (Laan, Polley, and Hubbard 2007), to show how such models can alleviate problems of inconsistency and lead to better release decisions.[2] Section 4 describes how a JBM's output

[1] These distinctions are at the heart of recent shifts in computer scientific thinking with respect to Artificial Intelligence. Take, for example, the development of Google Translate. As one author summarized: "the traditional symbolic AI approach to translating documents from one language to another proved ineffective. There are simply too many exceptions to the 'rules' governing how humans use languages for it to be practical to try to capture them all...What big data and statistical machine learning techniques have shown us is that given enough data many of these problems can be solved to a large degree, absent deep understanding, by looking for patterns in the data." http://watson.latech.edu/book/intelligence/intelligenceOverview5.html

[2] Our model is meant as proof of concept. Were administrators to actually implement a JBM, there would be no reason to restrict themselves to variables accessible from a hearing transcript: they would want to leverage the full set of data maintained by the CDCR.

could be adjusted to address the problem of systematic bias in decision making. Adjusting the output rather than the input allows for both optimal predictive performance and bias correction. However, unlike inconsistency, bias can only be corrected with a JBM if it is explicitly identified and estimated. Focusing on race in the criminal justice system, we discuss advances in techniques to identify and estimate bias. At the same time, we recognize that the challenge of bias estimation is considerable and describe a backup prophylactic approach that assures against an algorithm exacerbating existing bias. Section 5 considers how JBMs could be integrated into adjudication systems. We first discuss how JBMs should be understood in relation to the ascent of evidenced-based decision making in adjudication. We again focus on the criminal justice system, where recidivism forecasting is increasingly common and sophisticated. We note problems of relying on such algorithms – selection bias, policy distortion, the problem of measurement, and disparate racial impact – and the benefits of supplementing them with JBMs. We also consider challenges to implementing JBMs in the legal system: the tension between public transparency and variable manipulation, procedural fairness concerns, and issues of status quo bias. Section 6 concludes.

1 The Problems of Inconsistency and Bias

Complementing the long history of theoretical and qualitative accounts of inconsistency, researchers have more recently begun measuring it. In U.S. asylum cases, for example, research suggests at least 27% of cases would be decided differently if they were randomly assigned to a different judge (Fischman 2014). Chen et al. (2014) present evidence showing immigration judges are improperly influenced by previous decisions – they are up to 3.3% more likely to reject asylum if they granted asylum in their previous case. Matters are also bad at the appellate level, where it's estimated that roughly half of asylum appeals could have be decided differently had they been assigned to a different panel (Fischman 2014).

In a study of Israeli parole decisions, researchers showed that an inmates chances of parole decline precipitously the longer a judge has worked without a break (Danziger, Levav, and Avnaim-Pesso 2011). Abrams et al. (2012) find strong evidence of inter-judge disparities in the racial gap in sentencing in Cook County, Illinois.

Inconsistency, because it means we are failing to "treat like cases alike," is often thought to be a bad in itself. But the problem of inconsistency is not merely a problem of fairness. Inconsistency increases compliance and litigation costs by making it more difficult to anticipate the outcome of adjudication (Legomsky 2007). While scholars have noted the benefits of certainty in legal decisions (Kaplow 1992), the notion that inconsistency signals a sub-optimal system is less often acknowledged. We offer the following toy example to illustrate that inconsistency is incompatible with an optimal legal system. Assume that two commissioners, Commissioner H and Commissioner L, can both perfectly rank inmates in order of recidivism risk but Commissioner H is harsher and Commissioner L is more lenient toward inmates. Specifically, assume that that Commissioner H releases the safest 20% while Commissioner L releases the safest 40% of inmates. If those commissioners could be enticed to meet in the middle and both release the safest 30% – Commissioner L would exchange the release of inmates in the 30-40th percentile for Commissioner H's release of inmates in the 20-30th percentile – the compromise in the service of consistency would result in the release of a lower-risk parolee population. [3]

The problem of systematic bias in the legal system, in particular disparate racial treatment in the criminal justice system, is a subject of major concern. African Americans are vastly over represented in every stage of the criminal justice process, and the extent to which such disparities are a product of racial bias is a major subject of research. The majority of studies on the topic have been conducted using regression analysis, and the body of evidence

[3] Note that while inconsistency is a sign of a sub-optimal system, the inverse is not true: a system can be consistent but not optimal, as decision makers could consistently make poor decisions. Section 3 of the paper describes the conditions under which greater consistency results in better decisions.

from the research is inconclusive. More recent and sophisticated research suggests racial bias exists in at least some systems (Gelbach and Bushway 2011).[4]

2 Addressing Inconsistency with Judgmental Bootstrapping

2.1 Existing Literature on Judgmental Bootstrapping

The idea of judgmental bootstrapping, conceived in the early 1900s in relation to predicting corn crop quality, developed in a variety of fields in the 1960s (Dawes 1971). Dawes coined the term "bootstrapping" in his review of this research.[5] The central insight is that the fitted values from a regression model of expert judgments, by eliminating the uninformative variance or noise, will often be more highly correlated with the outcome variable being predicted than are the actual expert judgments themselves. Scholars in disciplines including psychology, education, marketing, and finance have applied judgmental bootstrapping to contexts ranging from school admissions decisions (Dawes 1971), predicting loan defaults (Abdel-Khalik and El-Sheshai 1980), criminal sentencing appeals decisions (Simester and Brodie 1993), and forecasting the number of advertising pages a magazine will sell (Ashton, Ashton, and Davis 1994). The method of modeling decisions to assist future decision-making has also been employed in settings outside of the academy: used by sports teams in making draft picks, for example, by corporations making business decisions, and in contemporary contexts such as the development of "robo-reader."

Evaluations of judgmental bootstrapping have demonstrated that bootstrapping gen-

[4] But others have found no evidence for racial prejudice on the part of parole boards (Mechoulan and Sahuguet 2015).

[5] Judgmental Bootstrapping is also sometimes referred to as "policy capturing" (Armstrong 2006).

erally improves the accuracy of expert forecasts (Armstrong 2006). In the one study cited in which judgmental bootstrapping failed to improve decision accuracy, it was shown to be the result of experts employing incorrect rules (Ganzach, Kluger, and Klayman 2000). We elaborate on the dangers of such backfiring JBMs in the following section.

2.2 A New Conceptual Framework: Crowdsourcing and Condorcet's Jury Theorem

We offer Condorcet's Jury Theorem as a new conceptual framework for understanding the effectiveness of judgmental bootstrapping.[6] The central idea of the classic political science theorem is that a group operating under majority rule is more likely to make an accurate decision than any random member of the group deciding alone. In the classic form, this is shown to hold true so long as (a) each individual's probability of making the right decision is greater than 50% and (b) the group members' votes are cast independently of one another. The requirement that all individuals must have a probability of making the right decision greater than .5 can be relaxed considerably: one only need to assume that the average of the individuals' probabilities is greater than .5 for the central insight of Condorcet's Jury Theorem to apply (Dietrich 2008). For simplicity, we present the weaker theorem shown formally with the following simple equation for the probability that N individuals each with probability p of making the right decision collectively arrive at the correct decision:

$$P(N,p) = \sum_{k=N/2}^{N} \binom{N}{k} p^k (1-p)^{N-k}$$

As the formula makes clear, the probability that a group will make an accurate decision

[6] Scholars have heretofore understood judgmental bootstrapping's efficacy as coming from its ability to infer "robust decision rules...which are not subject to the inconsistencies which can occur in human decision patterns due to various factors such as fatigue or distraction" (Lafond et al. 2015).

increases as the size of the group increases. Intuitively, if votes tend to be correct, then more votes are better because any one vote might be randomly mistaken.

Judgmental bootstrapping leverages this central insight. By modeling across and within judges, a JBM attempts to simulate a world in which (1) each judge votes in each case, to address inter-judge inconsistency, and (2) a judge casts multiple votes for each case under different circumstances (e.g. different times of day, in different moods, following different decision streaks), to address intra-judge inconsistencies. So, for example, a perfect JBM that generates a predicted probability of .63 for a particular parole case is indicating that, were these extensive set of hypothetical votes to be cast, 63% of votes would be in favor of release. In almost tautological phrasing, so long as those hypothetical votes would be correct more often than not, the perfect JBM coupled with a majority rule threshold (i.e., grant parole if the predicted probability is greater than .50) becomes increasingly likely to generate a correct decision as it accurately simulates more votes.

Judgmental bootstrapping can thus be effective if decision-makers tend to make good decisions (i.e., if votes tend to be correct) and the model can accurately simulate those votes. By averaging out deviations in judgment that stem from random influences – such as the idiosyncrasies of the judge assigned to a case, whether the case happens to be scheduled before or after a judges lunch break, or whether the judge happened to grant parole in the three cases preceding an individual's case – JBMs can boost the signal of the good judgments, isolating those judgments from the contamination of the random influences.

At the same time, just as JBMs will improve outcomes if decision-makers tend to make good decisions, aggregation will also intensify the signal of the bad judgments if judges tend to make poor decisions (Kerr, MacCoun, and Kramer 1996; Kerr and Tindale 2011). In this context, random influences are actually desirable. Through randomness, judges will at least sometimes make good decisions. A JBM should thus only be introduced into a system if we believe that judges are, on average, making good decisions.

However, even in cases in which judges tend to make good decisions overall, the problem of bias is not necessarily resolved. Specifically, judges might still tend to make poor decisions in a set of cases we are particularly concerned about. For example, given the history and evidence of racial bias and concerns about disparate racial impact in the criminal justice system, there is particular reason to worry about boosting racial biases. If this is case, it will be of little solace that aggregation is working well overall. However, as we explain in the Section 4, there are methods to both guard against boosting bias and to reduce pre-existing bias.

2.3 Machine-Learning JBMs

The traditional bootstrapping method "involves developing a model of an expert by regressing his forecasts against the information that he used" in order to infer the rules that the expert is using (Armstrong 2006). This approach is often effective, and offers the advantage of interpretability. But in contexts where decision-making requires nuanced judgments, its shortcomings can be stark. Most importantly, often we have no ability to measure much of the information that experts use in a complex decision task, such as most legal decisions. Without that information, a traditional JBM will, along with removing noise, remove much of the signal that it is designed to capture.

We advocate using a machine-learning approach to judgmental bootstrapping that builds a flexible, nonparametric model of decisions rather than a linear regression model built to infer decision rules. Importantly, we dispense with the restriction that the model includes only variables used by the expert. Instead, a machine-learning approach leverages variables that are merely correlated with the signal in order to capture it, so we include any variables that might be helpful in predicting decisions. The only exceptions are the noise variables – those variables that we have good reason to believe are statistically unrelated to

Table 1: Why Machine Learning?

Traditional Regression Approach	Machine-Learning Approach
Needs to model the experts' decision rules	Only needs to model the experts' decisions
Tries to infer decision rules	Exploits mere correlation
Hand-selects important variables	Lets the data determine the variables
Includes only those variables "used" by the expert	Only excludes variables thought to be randomly distributed with respect to case merits
Requires parametric assumptions	Is non-parametric
Danger of embedding modeler's normative preferences in the model	Separates normative preferences from the predictive model
Aggressively tries to reduce noise	Aggressively tries to capture signal
More likely to reduce bias	May boost biases if left unaddressed

the actual merits of a case. These variables may include, for example, the judge to which one happens to be assigned, the results of the immediately preceding cases, the time of day, whether the judge's football team won the night before, the weather, and the judge's mood.

A machine learning approach also has the benefit of separating the predictive task from controversial normative choices. The traditional approach puts too many choices in the hands of the modeler. Different models will generate different recommendations, and a modeler may, intentionally or not, choose a model that generates recommendations in accordance with her own normative preferences. A machine learning approach lets the data determine which model or combination of models is the best predictor.

3 Illustrating Judgmental Bootstrapping with California Parole Decisions

In the following section we analyze California Board of Parole suitability hearings to illustrate the potential of JBMs to improve decision-making. We present the Super Learner JBM, and provide evidence that such a model, in reducing inconsistency, could lead to better decisions.

The determination of parole suitability is made by the Board of Parole Hearings ("the Board"), an executive branch agency within the California Department of Corrections and Rehabilitation (CDCR), and release decisions are reviewed by the Governor. Through a public records request we obtained the population of California Board of Parole suitability hearing transcripts conducted since 2009, the first year in which the universe of transcripts are available electronically.[7]

The dataset was built using Python regular expressions to pull key information from each hearing transcript. The extracted information consists of: the commitment crime, the psychological risk assessment score as well as the identity of the evaluating psychologist, the minimum eligible parole date, the inmate's lawyer, the district attorney if present at the hearing, the number of victims present at the hearing, whether or not an interpreter was present at the hearing, the results of any previous suitability hearings, the inmate's date of entry into prison, and information concerning how many times the inmate has appeared before the Board, and the inmate's prison. We also extracted information on a limited number of noise variables: the presiding and deputy commissioners, the date and time of the hearing, and the results of the immediately preceding hearings.

Our JBM is meant as proof of concept. The model is limited to the variables that we could accurately extract from hearing transcripts. If administrators of the parole system were to actually implement a JBM, it should be built with the more expansive set of variables maintained by the CDCR and the Board of Parole Hearings.

[7] Although we have data going back to 2009, the paper restricts the analysis to 2011-2014 because of the stark difference in parole board practice since Governor Brown took office. The only exception is that for purposes of constructing the JBM (but not evaluating) we make use of all years.

3.1 Building and Describing the JBM

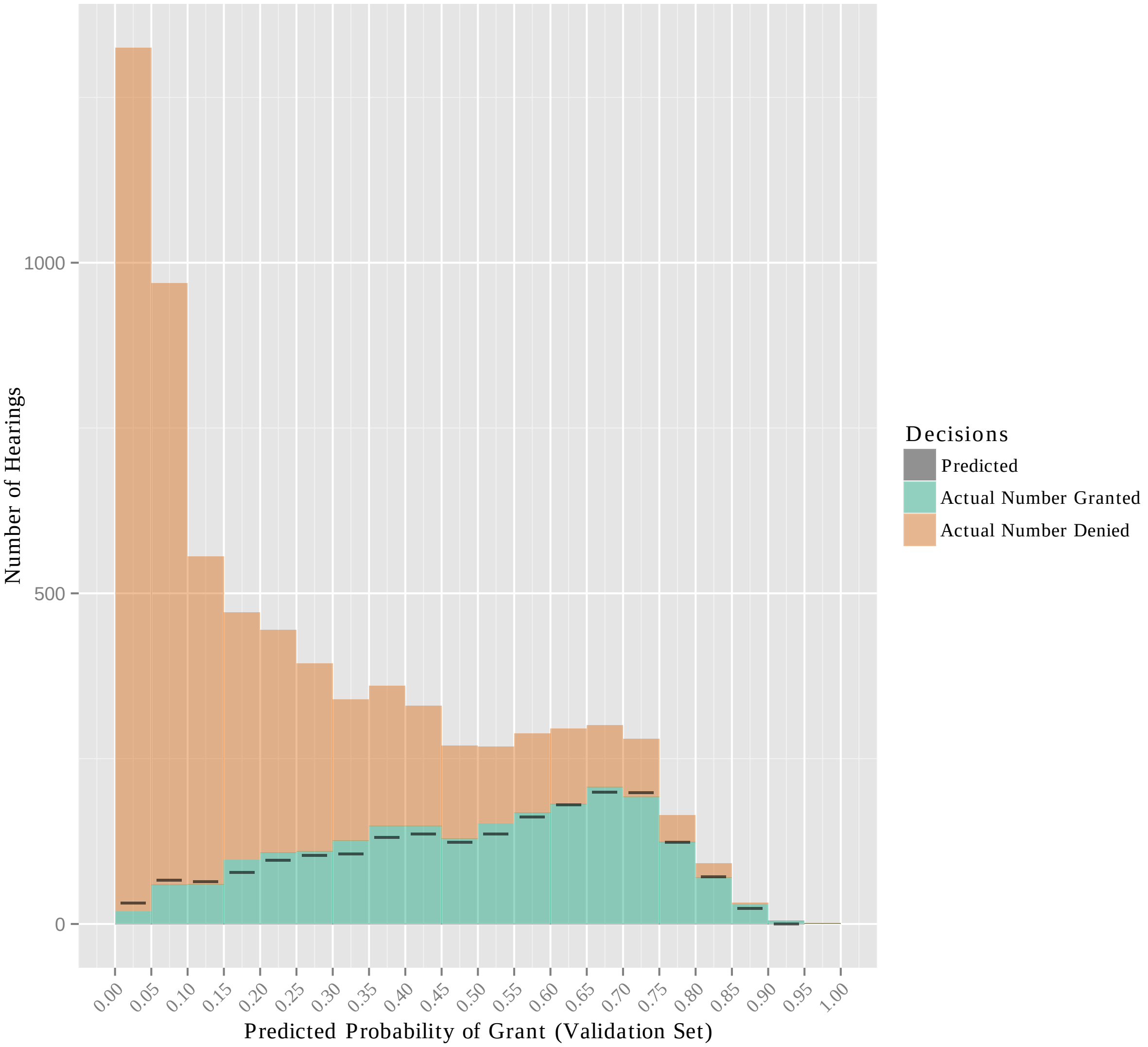
In building the JBM, we use only variables that are available pre-hearing and exclude the noise variables. We do not use information about the verbal exchanges that occur during the hearing. While text analysis of inmate speech could only increase predictive power of the model, inmate speech may partly reflect the commissioners' idiosyncrasies or inclinations to grant or deny parole. For example, an inmate who said "no" repeatedly might be facing questions from a tough commissioner or a commissioner who is otherwise in the mood to deny parole to the inmate. Using text analysis to help inform predictions might therefore give commissioner idiosyncrasies – arbitrary elements of the system that a JBM is designed to eliminate – an improper role in the model.

We construct our predictive algorithm using Super Learner, an ensemble machine-learning method developed in University of California Berkeley's Biostatistics department (Laan, Polley, and Hubbard 2007). Super Learner takes as input any number of user-supplied models (e.g., a parametric linear regression, random forest, lasso, etc) and combines those models' predictions to generate "super" predictions. Specifically, the Super Learner proceeds in two steps: first, validation-set predictions are generated for each candidate model; second, the true outcome is regressed on the candidate models' predictions to assign each model's predictions a weight. We offer a more extensive discussion of Super Learner generally, and our model specifically, in the appendix.

3.1.1 JBM Predictions

Figure 1 offers a graphic representation of the model's predictions. It shows the actual number of hearings that resulted in a grant or denial against the model's predicted probabilities. Figure 1 reveals the wide distribution of accurate predicted probabilities provided by our model. It also shows the particularly robust performance of our model in identifying hearings that have a very low probability of resulting in a grant. Note that an uninformative model would have simply one bar over the unconditional mean; a perfectly predictive model would have one red bar at zero and one green bar at one.

Figure 1: Validation Set Parole Predictions: 2011 - 2014



Our model correctly predicts validation-set 2011-2014 suitability hearing decisions with 79% accuracy. While correct classification rates have been the primary metric by which recent high-profile efforts to predict legal decisions, the Supreme Court's in particular, have been assessed (Ruger et al. 2004; Katz, Bommarito, and Blackman 2014), model performance is better evaluated with outcome probabilities. Using the Area Under the Receiver Operator Curve (the "AUC"), a measure of model performance that goes beyond classification accuracy, our model also provides valuable information with respect to the *probability* that an inmate is granted parole. The AUC provides the probability that a randomly selected case with a positive result (a hearing that ended in a grant) would have a higher predicted probability than a randomly selected case with a negative result (a hearing that ended in a denial). A model with an AUC score of 1 can perfectly discriminate between grants and denials; a model that is no better than random would score a 0.5. Our model has a cross-validated AUC of .80. This means there is an 80% probability that a randomly selected hearing that results in a grant was given a higher predicted probability than a randomly selected hearing that resulted in a denial. It is important to note, however, that a JBM should not be a perfectly predictive model. For a JBM to contribute value, it must explain less than 100% of the decisions. The entire point of the JBM is to eliminate the influence of factors that, although are related to decisions, are not related to the actual merits. Section 3.2.2 includes evidence that the signal capture is good enough to help the Board make better decisions – even though the model is handicapped by restriction to variables retrievable from hearing transcripts.

3.2 Using the JBM to Address Inconsistency

In this section we describe how and why a model of commissioner decision making could be used to help commissioners make better release decisions by reducing inter and intra commissioner inconsistency. We begin with details on how the JBM might actually be incorporated into a system.

3.2.1 Implementation Details

There are multiple ways an algorithmic prediction of a judgment could be incorporated to inform that judgment, with implementation differing along at least two important dimensions. First, an algorithmic recommendation could be more or less binding. For example, the algorithmic prediction could be offered as a mere recommendation, serving as an anchor, but with commissioners free to ignore it. Alternatively, on the more prescriptive end of the spectrum, the prediction could be used as a third vote on a two-member panel, such that a single decision maker's agreement with the algorithm would be sufficient to decide the outcome. Second, the algorithmic recommendation could be more or less granular in its presentation of the model's predicted probabilities to the decision makers. It could, for example, be presented as a binary recommendation: if the predicted probability of an inmate's parole is above a specified threshold, the algorithm recommends parole. In its most granular form, the recommendation would simply be the predicted probability. And a compromise might be the presentation of both the threshold recommendation as well as the distance between the inmate's predicted probability and threshold. Regardless of the exact implementation, the notion is that the commissioner decisions would remain discretionary but would coincide more frequently with the algorithm than they would in the absence of the implementation.

The advantages and disadvantages of different implementations are both under-theorized and empirically untested, and we do not attempt to determine which implementations would work best. Nonetheless, a recommendation threshold is likely a critical component of a successful implementation and deserves additional discussion. With complete granularity comes reduced ability to mitigate inconsistency. Although predicted probabilities could help commissioners control their own inconsistencies, the absence of a recommendation threshold would undermine the JBM's ability to address inter-commissioner inconsistencies: commissioners are likely to treat predicted probabilities in different ways (e.g., a lenient commissioner might consistently release more inmates with lower predicted probabilities). Thus, a recommendation threshold is likely to be essential, which then begs the question: what should the threshold be?

The recommendation threshold need not, and in fact often should not, be set at .5 (majority rule). The reason is twofold. First, while aggregation aims to synthetically implement majority rule to leverage the power of Condorcet's Jury Theorem, it will likely come up short in that effort. We cannot create a JBM that will perfectly predict the percentage of commissioner votes that would be in favor of release. Specifically, when the base-rate is low (e.g, the overall percentage of parole grants is significantly lower than .5), a JBM will often struggle to generate predicted probabilities above .5 (cite). Lowering the threshold can help account for an imperfect JBM. Second, even if a JBM were capable of perfectly representing vote share, a recommendation threshold of .5, which would effectively implement majority rule, might disturb existing political compromises. For example, at the extreme, we might imagine an actual system of majority rule that would lead to no parole releases because all actual parole releases rely on high levels of noise in the system (such as lenient commissioners getting to make decisions by themselves). These occasional, noise-induced releases could reflect a political equilibrium that a JBM should attempt to respect.

In light of these concerns, a good default recommendation threshold is one that respects the general values of the entire decision-making system. We thus suggest that the default threshold is one set a level that, if applied to recent historical data, would recommend release for the same number of inmates as were actually released. A JBM, if it's recommendations were followed, can therefore result in release of better inmates holding constant the number of inmates to whom the judges grant parole. Notably, this result abstracts away from contentious debates over whether parole judges are releasing too many or too few inmates and holds regardless of the rate of release. Accordingly, we hypothetically set a recommendation threshold of .43 for our JBM. From 2011 to 2014, the Board granted parole to approximately 30% of inmates, and a threshold of .43 would have resulted in grant recommendations to approximately 30% of the inmate population.

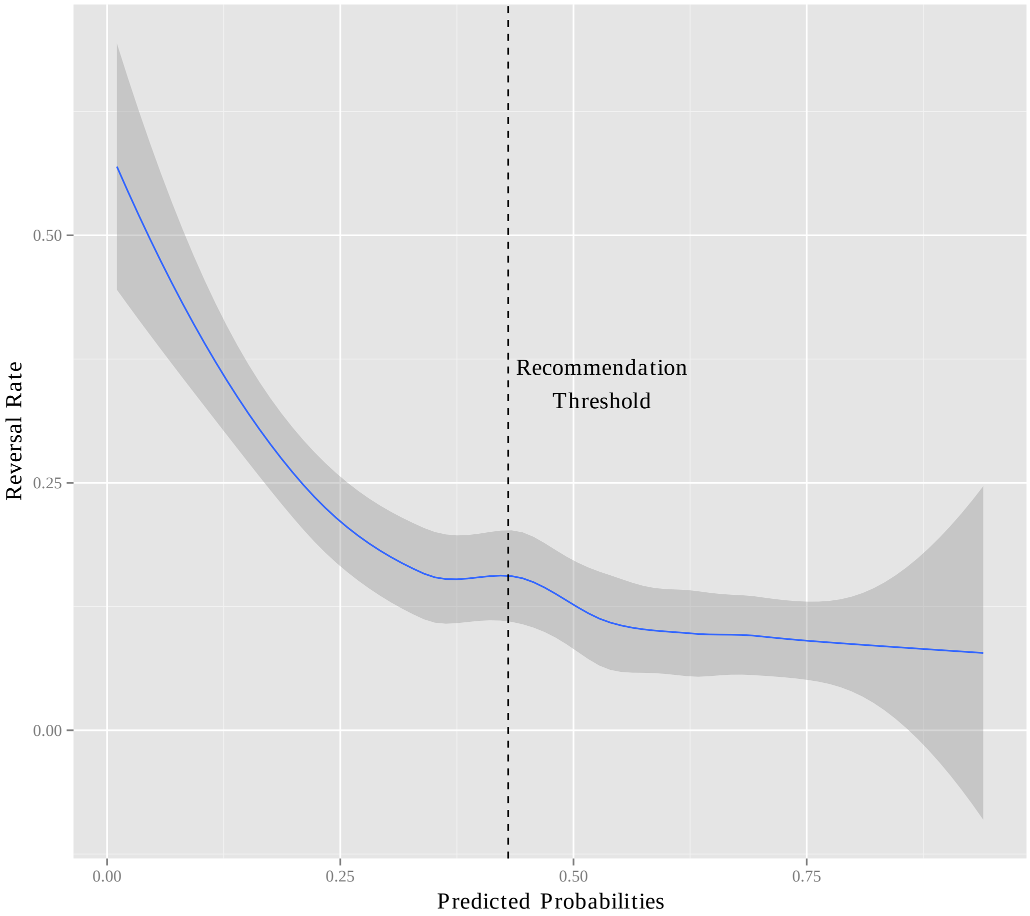
While a threshold that targets recent historical values is a reasonable default, particularly from a political buy-in perspective, the threshold can be adjusted to reflect new values. For example, assume it's the Governor's office that is implementing a JBM in order to help his appointee's make better decisions. If the Governor thought it prudent to release more parolees, lowering the threshold provides an appealing and principled way to accomplish the goal. He need not wait to find new, qualified appointees who share his values – he can use a JBM to distill the skilled judgment of the Board but then adjust the recommendation threshold to make it more lenient.

3.2.2 Evidence of Better Release Decisions

We cannot directly test whether our JBM would produce better release decisions. There is, of course, no exact measure of "better," and we do not have access to a measure of re-offending, a natural proxy for it. But we think Governor Jerry Brown's reversal decisions provide evidence that the JBM recommendations could improve commissioner decision-making. Governor Brown has the ability to reverse any parole granted to an inmate convicted of murder. Figure 2 shows relationship between validation set predicted probabilities and the Governor's reversal rate.[8] If our JBM recommended against parole, the Governor was two and a half times more likely to reverse the Board's decision.

[8] As a proxy for governor reversals, we denote a case as reversed if an inmate's parole is granted but the inmate reappears in another hearing at a later date. In order to avoid problems of potentially biased missingness, we restrict the analysis to 2011 and 2012 hearings so that there is sufficient passage of time for inmates to show up in the dataset again if their parole is reversed.

Figure 2: Relationship Between Model Predictions and Governor Reversal Rate (Loess Smoothed)



4 Adjusting the JBM to Address Bias

While Figure 2 provides evidence that the JBM is functioning well overall, it does not address concerns about systematic biases. Perhaps counterintuitively, the problem of bias cannot be resolved by the exclusion of the variables that we think judges may exhibit bias over. For example, if we were concerned about racial bias in parole decisions, excluding race from the JBM would not adequately address the concern because other variables may serve as proxies for race, and the JBM could then simply capture racial biases through those proxies. Attempting to eliminate the proxies for race is also unlikely to be successful. Finding these proxies is difficult: ostensibly benign variables might interact in complex ways to relate to race. Additionally, the process of removing variables can significantly reduce the JBMs ability to boost the good signal: a variable that might be related to race might also be related to good judgments, so excluding it would be a version of "throwing the baby out with the bathwater."

We instead propose addressing bias directly by adjusting the algorithm's output. We recommend estimating bias and then explicitly correcting for it by adjusting the algorithmic recommendations. For example, if the estimate of disparate treatment of black offenders is a 10% reduction in parole chances, this method would add 10% to black offenders' predicted probabilities.

The challenge of this approach is the challenge of estimating biases with observational data. However, recent advances in data collection and estimation techniques are allowing for more plausible causal inference. Methods such as propensity score matching (Hirano, Imbens, and Ridder 2003; Abadie and Imbens 2006) and those that combine treatment modeling with outcome modeling (Rubin 1979; Van der Laan and Robins 2003) provide extra traction in observational studies. And these advances in research design are increasingly being coupled with non-parametric, data-adaptive estimation techniques, such as genetic matching (Diamond and Sekhon 2013) and targeted learning (Van der Laan and Rose 2011). These new design and estimation techniques have the additional benefit of deescalating the war of experts wherein one expert offers a model that (predictably) shows bias and another offers a model that (predictably) doesn't. By separating the research design from the estimation procedure, the methods can guard against specification searching. And while data-adaptive estimation techniques cannot eliminate the need for assessing which variables must be controlled for, they can at least largely eliminate the debate over *how* variables are controlled for by allowing the data to determine which models are best.

These advances in data, design and estimation are unlikely to ease all concerns of omitted variable bias, but the threshold of certainty required for the advancement of scientific knowledge should often be relaxed in policy world contexts that require action. Academics enjoy the luxury of waiting around or choosing questions based on the availability of a strong quasi-experimental research design, but the world sometimes demands a best answer. One might proceed even without robust causal estimates if the biases are particularly important to address.

At the same time, we acknowledge the difficulty of estimating the existence and degree of bias, and so also offer a prophylactic approach that merely aims to assure the algorithm does not inflate biases. The method uses the status quo as a safeguard. Specifically, recent decisions can be compared to what the JBM would have recommended and if relevant disparities are widened under the algorithmic recommendations, group-specific recommendation thresholds may be adjusted so as to maintain the status quo outcome distribution. For example, if a JBM recommends parole for 20% of black inmates but 25% were granted parole in recent history, this might be evidence the algorithm is inflating bias against black inmates. A separate recommendation threshold could be implemented to set the recommendation rate equal to the actual historical rate of 25%.: if a threshold of .35 would have generated parole recommendations for 25% of black inmates (the percentage that were historically released), then .35 would be the group-specific recommendation threshold for those inmates.[9]

[9] We note that both methods we present for addressing bias have a cost. If there is no bias, bias is poorly estimated, or there are highly heterogeneous treatment effects, then the adjustments can lead the JBM to generate inferior recommendations. Which approach is preferred will depend in large part on the confidence in the estimates of bias, but legal considerations might also influence the choice. The first method, despite being more interventionist and requiring difficult estimations, might be more likely to pass constitutional muster if applied to race. An estimate of bias, even if made without confidence, presents evidence of a constitutional problem that requires addressing. The prophylactic approach, on the other hand, has at least surface-level similarities to the type of mechanical race-based adjustments that have been found unconstitutional in the higher education context.

5 Additional Considerations

5.1 Integrating JBMs with the Rise of Evidence-Based Adjudication

Outside of the legal context, algorithms help address concerns about arbitrariness and bias in human decision making. In the medical domain, for example, algorithms predicting a patient's disease can help address the deficiencies of doctors' diagnoses: with a database of previous patients' symptoms and ultimate afflictions, can be used to generate a predictive model to help more accurately diagnose the patient. In the legal domain such algorithms are less prevalent; we often can't build an algorithm to predict the right legal outcome, as there is no agreed upon way to measure it.

There are some legal contexts in which we have access to a reasonable proxy for the "correctness" of a decision. The criminal justice system, for example, has long employed criminal risk assessments, and they are increasingly sophisticated. Richard Berk (2012; 2016), for example, has been advocating the use of machine-learning tools to predict re-offending likelihood.

While such proxy-based models hold great promise, perhaps even in legal contexts beyond criminal justice, there are deficiencies that JBMs can help address. We discuss three weaknesses with proxy-based models, with a focus on recidivism forecasting: selection bias, policy distortion, and problems of measurement.[10]

By virtue of the fact that legal decisions have important consequences for the fates of the legal system's users, the reliability of proxy-based models may be undermined by problems

[10] There are a number of ways in which a JBM could be incorporated with proxy-based algorithms. For example, a JBM could help assess the extent to which selection bias is indeed a real concern in a given proxy-based algorithm and proxy algorithms could be used to evaluate the quality of the JBM recommendations.

of selection bias. In the criminal justice system, for example, researchers and administrators have used measures of recidivism as an outcome that can allow for the construction of models to aid decision making. But the problem of selection bias for recidivism prediction, long recognized in the criminological literature (Smith and Paternoster 1990; Loughran et al. 2009), presents reason to worry about their accuracy. The crux of the problem is the mismatch between the dataset used to build a recidivism model and the set of individuals to whom the model is applied. As Berk et al. (2016) acknowledge, it is difficult to generalize forecasts of recidivism to the set of individuals who never had an opportunity to recidivate. For simplicity, consider models used to assist parole decisions, although the argument also applies to bail and sentencing decisions. The model is fit on a dataset of individuals granted parole – the group for which we have a recidivism measure – but the model is applied to all parole-eligible inmates. However, information about paroled individuals may not provide accurate forecasts for individuals that a judge would not have paroled. This is like fitting a model on apples to forecast both apples and oranges to help decide who should be an apple. Judges do not, presumably, release observably similar individuals randomly, so there is reason to worry about the application of forecasts of paroled inmates to the entire population of parole-eligible inmates. [11]

Even if the issue of selection bias can be overcome, proxy-based models may undesirably distort legal policy towards the measurable. As it is said, "Not everything that counts can be counted, and not everything that can be counted counts."[12] Recidivism models, for example, will encourage judges to focus on reducing recidivism at the expense of other objectives. As critics have argued, there is reason to believe judges will privilege the ostensibly scientific predictions from a recidivism model, leading to an increased emphasis on reducing recidivism

[11] It is important to distinguish between data-driven risk assessment instruments and psychological risk assessments. In the case of the psychological assessments, the assessment is theoretically grounded and constructed a priori. Because recidivism data is not used in the construction of the tool, but is instead used only to validate it, our selection-bias objection does not apply to psychological risk assessments.

[12] Though frequently attributed to Albert Einstein, a very similar version of the quotation can be found in William Bruce Cameron's 1963 text "Informal Sociology: A Casual Introduction to Sociological Thinking."

(Starr 2014). While recidivism may be of primary interest in many criminal justice systems, it is not the only consideration. Deterrence, rehabilitation, the negative consequences of continued incarceration on an individual's personal and family life, and satisfying retributive impulses may be relevant as well.

Furthermore, the availability of a readily measurable and objective proxy may be more illusory than real. We often have only a proxy for a proxy. With recidivism models, for example, the real interest is in future criminality. Yet this cannot be directly assessed. Instead, it is imperfectly measured with arrests, convictions, or prison return data. And the problem of what should count as recidivism - what type of future criminal activity we actually care about – has long plagued criminal justice system researchers and administrators. The recent Pennsylvania proposal for evidenced-based sentencing illuminates these problems of definition and measurement. The Pennsylvania Commission on Sentencing, in compliance with a law that requires it to develop a instrument to assist the court in sentencing, has begun developing predictive models of recidivism. In its current form, the Commission uses "re-arrest and, for offenders sentenced to state prison, re-incarceration *on a technical violation*" (Pennsylvania Sentencing Commission 2015) (emphasis added). Although reports indicate the Commission intends to analyze different types of recidivism risk (i.e., for violent and nonviolent crimes), in the eleven reports it has so far completed, they have only addressed the bluntest of questions: what are an individual's chances of recidivating in any way? Thus, while "recidivism" may be measurable, the choices over how it is measured reflect value judgments, and claims to objectivity should be treated with suspicion. This problem of measurement also dovetails with concerns about racial biases in the criminal justice system. If recidivism includes crimes that are subject to police discretion, instruments such as Pennsylvania's can compound racial disparities by producing artificially high risk scores for individuals in heavily policed and supervised minority communities. If, as evidence suggests (Beckett et al. 2005), black men are more likely to be caught for drug possession than their white counterparts, and "recidivism" includes re-arrest for drug possession, risk scores for black males will be erroneously inflated. Similarly, if black men are more frequently penalized for technical parole violations, a measure of recidivism that includes such violations will unfairly increase risk scores for black men.

By modeling decisions rather than a post-decision outcome of interest, JBMs avoid the above noted problems with proxy-based models. JBMs escape problems of selection bias by modeling all decisions rather than outcomes for a non-random subset of individuals, they provide recommendations that reflect a complex set of policy considerations rather than a single measurable proxy, and JBMs do not rely on controversial measurements of a proxy.

Despite the weaknesses of proxy-based models outlined above, they may still be useful, especially in conjunction with JBMs. First, the problems we have outlined are not insurmountable, at least as a theoretical matter. For example, Berk et al. (2016) propose testing for the problem of selection bias by having judges record a hypothetical decision before they see a recidivism prediction.[13]. Second, JBMs and proxy-based algorithms can be combined to offer a triangulated approach to making better legal decisions. If the algorithms correspond in their recommendations, a judge can feel more confident deferring to them. When the algorithm recommendations differ, it would suggest the need for greater judicial discretion.

5.2 Some Pragmatic Concerns

In this section, we note and address some of the concerns with implementing JBMs. Perhaps most obviously, judgmental bootstrapping inherently ties future decisions to the past, a past that might be incompatible with current values. In the California parole context, an algorithm built in the early 2000s would have helped to entrench an era in which almost no one was released. So while a JBM offers the advantage of producing temporal consistency and a version of institutional memory, it also supports the status quo and encourages a

[13] We also suggest leveraging inter-judge disparities to see test how well recidivism predictions generalize to the marginal release.

value-based path dependence. While this is a legitimate concern, we think JBMs can be a solution to inconsistency even in a world with rapidly changing values. First, despite being based on the past, JBMs also present a unique opportunity to quickly update a system in accordance with new values. A simple adjustment to the algorithm, such as lowering the release recommendation threshold or making a racial bias correction, could effect change that might otherwise come only with slow shifts in judicial attitudes. And even without direct intervention, a JBM can be updated as values shift. Insofar as judges express new values by refusing to follow JBM recommendations (generated from past decisions), datasets of the judges' decisions could be used to make the JBM compatible with current values.[14] Furthermore, proxies such as recidivism could be used to continuously test whether a JBM has become outdated. A final and, we think, potentially more serious concern in this vein is that a system that has come to rely on a JBM might then be reluctant to make changes to the system which would undermine the usefulness of the JBM. For example, the California Board of Parole Hearings recently altered the psychological risk assessment tool used to evaluate inmates: inmates are now scored as high, moderate, or low risk, whereas scores of high-moderate and low-moderate were previously also possible. The change in the assessment, which could conceivably improve decisions, also likely lowers the predictive capacity of the JBM. If the JBM had been instituted, it's imaginable that administrators might have been reluctant to make the change to the risk assessment tool. Of course, the reluctance to change could be good policy if the change is merely trivial or cosmetic. A JBM's inconsistency and bias reductions should be weighed against the value of a change.

A second concern with using JBMs in legal decision making is that such algorithmic procedures may conflict with fairness intuitions. Justice in sentencing, some argue, requires "an individualized assessment of the offender and the offense, leading to a moral judgment

[14] But what if despite shifting values that conflict with JBM recommendations, judges simply follow recommendations out of habit or ease? To address this concern, some random subset of cases could be chosen to proceed to a trial case without the aid of the JBM. Departures from simulated JBM recommendations could alert administrators to a malfunctioning JBM and be used to update the model.

imposed by judges with skill, experience, and wisdom" (Luna 2002). Social psychologists have studied and empirically verified the psychological costs of violating this sense that "individualized" treatment is required for fairness. This procedural fairness research shows people care greatly about the process by which outcomes are reached, even if the result is an outcome they find unfavorable (Lind & Tyler 1988). Insofar as people think equitable and legitimate decisions are those in which each person is treated individually with the nuances of each case addressed, an algorithm-assisted approach may violate this sense of procedural fairness. At the same time, it is not at all clear that fairness intuitions would be in greater conflict with a JBM-assisted process than they are in a system where the idiosyncrasies or mood of a particular judge can dramatically affect the outcome of a case.

Third, there is a tension between successfully implementing a judgmental bootstrapping model in a legal system and making information about the model public. Inasmuch as a model uses predictive variables that can be inexpensively manipulated, users of the legal system may be able to strategically alter their variables so as to obtain more favorable algorithmic recommendations. Consider, for example, an inmates attorney in the California parole system. Some private attorneys are moderately associated with higher chances of parole. While that may be in part causal, some of the association is likely due to correlation with unobservables: those more eligible for release may also be more likely to hire a private attorney. If inmates know that an algorithm will give them stronger recommendations if they hire certain attorneys, they may be more likely to do so – even if they aren't actually the type of inmate who is eligible for release. Similarly, inmate knowledge of an algorithm may alter the pool of inmates who actually proceed with their scheduled hearings. Inmates who in the absence of the algorithm would not have proceeded with their hearings – and thus would not have been part of the population on which the algorithm was built – might take advantage of an algorithm that, because they happen to share characteristics with parole-eligible inmates, generates positive recommendations.[15] And the problem extends to other

[15] This problem of selection bias could be at least partially addressed in legal systems which maintain

actors as well. District Attorneys or victims might also alter their behavior to take advantage of an algorithm. The fact that commissioner judgment would only be guided by algorithmic recommendations would mitigate these types of issues, but insofar as an algorithm does alter judgments (which is, of course, the entire point), public knowledge of an algorithm's inner workings can lead to problems of variable manipulation and selection bias. While removing easily manipulated or selection-inducing variables from the algorithm can mitigate the problem, it comes at the price of predictive performance.

6 Conclusion

We have suggested a novel two-step algorithmic approach to addressing inconsistency and bias in legal decision making. Specifically, judgmental bootstrapping models built with machine learning methods, by providing judges with recommendations generated from statistical models of themselves, can help those judges make better and more decisions. Acknowledging that decision-maker biases can be embedded in and even inflated by a JBM, we argue the issue is best addressed by explicit adjustments to the algorithm's output. Attempts at removing the bias-inducing input from the algorithm are both likely to be unsuccessful and to reduce predictive performance.

References

Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." 01383, *econometrica* 74 (1): 235–267. Accessed May 22, 2016. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0262.2006.00655.x/abstract.

data on the full population of potential users.

Abdel-Khalik, A. Rashad, and Kamal M. El-Sheshai. 1980. “Information Choice and Utilization in an Experiment on Default Prediction.” 00120, *Journal of Accounting Research:* 325–342. Accessed October 7, 2015. http://www.jstor.org/stable/2490581.

Abrams, David, Marianne Bertrand, and Sendhil Mullainathan. 2012. “Do Judges Vary in Their Treatment of Race?” 00113, *Journal of Legal Studies* 41 (2): 347–383. Accessed October 7, 2015. http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1800840.

Armstrong, J. Scott. 2006. “Findings from Evidence-Based Forecasting: Methods for Reducing Forecast Error.” 00148, *International Journal of Forecasting* 22 (3): 583–598. Accessed October 5, 2015. http://www.sciencedirect.com/science/article/pii/S0169207006000537.

Ashton, Alison Hubbard, Robert H. Ashton, and Mary N. Davis. 1994. “White-Collar Robotics: Levering Managerial Decision Making.” 00009, *California Management Review* 37 (1): 83. Accessed October 7, 2015. http://search.proquest.com/openview/8590459ba0156326aa9c3ad8
1?pq-origsite=gscholar.

Beckett, Katherine, Kris Nyrop, Lori Pfingst, and Melissa Bowen. 2005. “Drug use, Drug Possession Arrests, and the Question of Race: Lessons from Seattle.” 00131, *SOCIAL PROBLEMS-NEW YORK-* 52 (3): 419. Accessed October 5, 2015. http://socpro.oxfordjournals.org/content/socpro/52/3/419.full.pdf.

Berk, Richard. 2012. *Criminal Justice Forecasts of Risk: A Machine Learning Approach.* 00021. Springer Science & Business Media. Accessed May 22, 2016. https://books.google.com/books?hl=en&lr=&id=Jrlb6Or8YisC&oi=fnd&pg=PR3&dq=Richard+Berk+(2012+machine+learning&ots=IuCe6dipsc&sig=CBHsKVxRWYVmFX4kxx8WOyFSyG0.

Berk, Richard A., Susan B. Sorenson, and Geoffrey Barnes. 2016. "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions." 00001, *Journal of Empirical Legal Studies* 13 (1): 94–115. Accessed May 22, 2016. `http://onlinelibrary.wiley.com/doi/10.1111/jels.12098/full`.

Chen, D. L., and H. Spamann. 2014. *This Morning's Breakfast, Last Night's Game: Detecting Extraneous Factors in Judging.* Technical report. 00009. Working paper, ETH Zurich.

Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue. 2014. "Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." 00012, *Loan Officers, and Baseball Umpires (November 4, 2014).* Accessed October 5, 2015. `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2538147`.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous Factors in Judicial Decisions." 00356, *Proceedings of the National Academy of Sciences* 108 (17): 6889–6892. Accessed October 5, 2015. `http://www.pnas.org/content/108/17/6889.short`.

Dawes, Robyn M. 1971. "A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making." 00537, *American psychologist* 26 (2): 180. Accessed October 7, 2015. `http://psycnet.apa.org/journals/amp/26/2/180/`.

Diamond, Alexis, and Jasjeet S. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." 00458, *Review of Economics and Statistics* 95 (3): 932–945. Accessed May 22, 2016. `http://www.mitpressjournals.org/doi/abs/10.1162/REST_a_00318`.

Dietrich, Franz. 2008. "The Premises of Condorcet's Jury Theorem are not Simultaneously Justified." 00039, *Episteme* 5 (01): 56–73. Accessed October 5, 2015. `http://journals.cambridge.org/abstract_S1742360000000927`.

Fischman, Joshua B. 2014. "Measuring Inconsistency, Indeterminacy, and Error in Adjudication." 00011, *American law and economics review* 16 (1): 40–85. Accessed October 5, 2015. http://aler.oxfordjournals.org/content/16/1/40.short.

Ganzach, Yoav, Avraham N. Kluger, and Nimrod Klayman. 2000. "Making Decisions from an Interview: Expert Measurement and Mechanical Combination." 00046, *Personnel Psychology* 53 (1): 1–20. Accessed October 7, 2015. http://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.2000.tb00191.x/full.

Gelbach, Jonah B., and Shawn D. Bushway. 2011. "Testing for Racial Discrimination in Bail Setting Using Nonparametric Estimation of a Parametric Model." 00000, *Available at SSRN 1990324.* Accessed May 22, 2016. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1990324.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." 01273, *Econometrica* 71 (4): 1161–1189. Accessed May 22, 2016. http://onlinelibrary.wiley.com/doi/10.1111/1468-0262.00442/abstract.

Kaplow, Louis. 1992. "Rules versus Standards: An Economic Analysis." 01831, *Duke Law Journal:* 557–629. Accessed October 9, 2015. http://www.jstor.org/stable/1372840.

Katz, Daniel Martin, Michael James Bommarito, and Josh Blackman. 2014. "Predicting the Behavior of the Supreme Court of the United States: A General Approach." 00012, *Available at SSRN 2463244.* Accessed October 5, 2015. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2463244.

Kerr, Norbert L., Robert J. MacCoun, and Geoffrey P. Kramer. 1996. "Bias in Judgment: Comparing Individuals and Groups." 00504, *Psychological review* 103 (4): 687. Accessed May 22, 2016. http://psycnet.apa.org/?fa=main.doiLanding&doi=10.1037/0033-295X.103.4.687.

Kerr, Norbert L., and R. Scott Tindale. 2011. “Group-based Forecasting?: A Social Psychological Analysis.” 00043, *International Journal of Forecasting* 27 (1): 14–40. Accessed October 5, 2015. http://www.sciencedirect.com/science/article/pii/S0169207010000166.

Laan, Mark J. van der, Eric C. Polley, and Alan E. Hubbard. 2007. “Super Learner.” 00267, *Statistical applications in genetics and molecular biology* 6 (1). Accessed October 5, 2015. http://www.degruyter.com/view/j/sagmb.2007.6.1/sagmb.2007.6.1.1309/sagmb.2007.6.1.1309.xml.

Legomsky, Stephen H. 2007. “Learning to Live with Unequal Justice: Asylum and the Limits to Consistency.” 00076, *Stanford Law Review:* 413–474. Accessed May 22, 2016. http://www.jstor.org/stable/40040413.

Loughran, Thomas A., Edward P. Mulvey, Carol A. Schubert, Jeffrey Fagan, Alex R. Piquero, and Sandra H. Losoya. 2009. “Estimating a Dose-Response Relationship Between Length of Stay and Future Recidivism in Serious Juvenile Offenders*.” 00000, *Criminology* 47 (3): 699–740. Accessed October 7, 2015. http://onlinelibrary.wiley.com/doi/10.1111/j.1745-9125.2009.00165.x/abstract.

Luna, Erik. 2002. *Misguided Guidelines: A Critique of Federal Sentencing.* 00004. Cato Institute. Accessed May 22, 2016. http://www.cato.org/publications/policy-analysis/misguided-guidelines-critique-federal-sentencing.

Mechoulan, Stephane, and Nicolas Sahuguet. 2015. “Assessing Racial Disparities in Parole Release.” 00001, *The Journal of Legal Studies* 44 (1): 39–74. Accessed May 22, 2016. http://www.jstor.org/stable/10.1086/680988.

Nakosteen, Robert, and Michael Zimmer. 2014. “Approval of Social Security Disability Appeals: Analysis of Judges’ Decisions.” 00001, *Applied Economics* 46 (23): 2783–2791. Accessed October 5, 2015. http://www.tandfonline.com/doi/abs/10.1080/00036846.2014.914147.

Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Philip G. Schrag. 2007. “Refugee Roulette: Disparities in Asylum Adjudication.” 00355, *Stanford Law Review:* 295–411. Accessed October 5, 2015. http://www.jstor.org/stable/40040412.

Rubin, Donald B. 1979. “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies.” 00671, *Journal of the American Statistical Association* 74 (366a): 318–328. Accessed May 22, 2016. http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1979.10482513.

Ruger, Theodore W., Pauline T. Kim, Andrew D. Martin, and Kevin M. Quinn. 2004. “The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking.” 00226, *Columbia Law Review:* 1150–1210. Accessed October 5, 2015. http://www.jstor.org/stable/4099370.

Simester, Duncan I., and Roderick J. Brodie. 1993. “Forecasting Criminal Sentencing Decisions.” 00003, *International Journal of Forecasting* 9 (1): 49–60. Accessed October 7, 2015. http://www.sciencedirect.com/science/article/pii/016920709390053P.

Smith, Douglas A., and Raymond Paternoster. 1990. “Formal Processing and Future Delinquency: Deviance Amplification as Selection Artifact.” 00096, *Law and Society Review:* 1109–1131. Accessed October 7, 2015. http://www.jstor.org/stable/3053663.

Starr, Sonja B. 2014. “Evidence-Based Sentencing and the Scientific Rationalization of Discrimination.” 00053, *Stan. L. Rev.* 66:803. Accessed October 5, 2015. http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/stflr66&section=24.

Van der Laan, Mark J., and James M. Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality.* 00609. Springer Science & Business Media. Accessed May 22, 2016. https://books.google.com/books?hl=en&lr=&id=z4_-dXslTyYC&oi=fnd&pg=PR5&dq=Unified+Methods+for+Censored+Longitudinal+Data+and+Causality.&ots=uPP9VEFqeN&sig=_cQyRe35av79OveVZmJteVc0eAA.

Van der Laan, Mark J., and Sherri Rose. 2011. *Targeted learning: causal inference for observational and experimental data.* 00216. Springer Science & Business Media. Accessed October 5, 2015. https://books.google.com/books?hl=en&lr=&id=RGnSX5aCAgQC&oi=fnd&pg=PR3&dq=Targeted+Learning:+Causal+Inference+for+Observational+and+Experimental+Data&ots=FPb9cuAT7B&sig=P-RCh8efs-QtJYp1duNsq2W7Tbw.

A Appendix

A.1 Super Learner

In order to generate validation-set predictions, Super learner breaks whatever data it is given into ten separate random "chunks."[16] The first chunk, the first 1/10th of the data, is then set aside and the underlying models are built using the remaining 9/10ths of the data. The left-out 1/10th of the data, the "validation set," is then plugged into the underlying models and used to generate model predictions. The same process is repeated for each of the remaining chunks. That is, the second 1/10th chunk of data is set aside, and Super Learner builds the models on the remaining 9/10 of the data (the first chunk is now being used to help build the model) and then generates validation set predictions for the second 1/10th chunk of data. And so on for all ten chunks. The appeal of these validation set predictions is that they allow us to estimate how the underlying model would perform on data it has never seen.

The first step generates validation set predictions for each data point for each underlying model. In the second step, Super Learner then leverages the cross-validation information on model performance to assign weights to each model according to how well their predictions match the true outcome. It does this by regressing the true outcome on the underlying model predictions.[17]

Our predictive model of the Parole Board decisions consists of fourteen candidate algorithms. We cross-validate the candidate algorithms as well as the Super Learner itself. In all cases, the Super Learner performs at least marginally better than any of the individual underlying models. Table 3 displays statistics for cross-validated mean squared error and

[16] Ten-fold cross-validation is the default. Users may choose other fold numbers, but ten-fold cross-validation is generally regarded as an appropriate choice.

[17] As a default, Super Learner runs a non-negative least squares regression.

Table 3: Cross-Validated Performance of Super Learner and Candidate Models

Model	Avg MSE	Min MSE	Max MSE	Weight
Super Learner: weighted composite of the below candidate models	0.134	0.124	0.143	NA
The overall average grant rate	0.187	0.172	0.200	0.002
Logistic regression with all variables	0.136	0.125	0.146	0.049
Logistic regression with the 33% of variables with smallest t-test p-vals	0.136	0.126	0.146	0.079
Logistic regression with the 50% of variables with smallest t-test p-vals	0.136	0.126	0.148	0.079
Logistic regression with the 66% of variables with smallest t-test p-vals	0.135	0.124	0.146	0.010
Logistic regression with interaction of the 6% of variables with smallest t-test p-vals	0.144	0.137	0.151	0.040
Classification and regression tree	0.149	0.140	0.160	0.000
Random Forest with default parameters (mtry = 33% of variables)	0.142	0.133	0.152	0.012
Random Forest with mtry = 10	0.138	0.130	0.147	0.043
Random Forest with mtry = 20	0.139	0.131	0.147	0.198
Bayesian generalized linear model with default parameters	0.136	0.125	0.146	0.198
LASSO regression	0.135	0.124	0.144	0.038
Elastic-net regularized generalized linear model with alpha=.1	0.135	0.125	0.145	0.021
Stepwise model selection by AIC	0.136	0.125	0.146	0.232

ensemble weights for the models. Three of the fourteen candidate models – a Random Forest, a Bayesian linear model, and a stepwise AIC-based model selector – account for more than 60% of the Super Leaner.