Almost Politically Acceptable Criminal Justice Risk Assessment

Richard Berk[1,2] and Ayya A. Elzarka[3]

[1]Department of Criminology, University of Pennsylvania

[2]Department of Statistics, University of Pennsylvania

[3]Google LLC, Mountain View, CA.

December 31, 2019

Abstract

Research Summary

In criminal justice risk forecasting, one can prove that it is impossible to optimize accuracy and fairness at the same time. One can also prove that usually it is impossible optimize simultaneously all of the usual group definitions of fairness. In policy settings, one necessarily is left with tradeoffs about which many stakeholders will adamantly disagree. The result is a contentious stalemate. In this paper, we offer a different approach. We do not seek perfectly accurate and perfectly fair risk assessments. We seek politically acceptable

risk assessments. We describe and apply a machine learning approach that addresses many of the most visible claims of "racial bias" to arraignment data on 300,000 offenders. Regardless of whether such claims are true, we adjust our procedures to compensate. We train the algorithm on White offenders only and compute risk with test data separately for White offenders and Black offenders. Thus, the fitted, algorithm structure is exactly the same for both groups; the algorithm treats all offenders as if they are White. But because White and Black offenders can bring different predictors distributions to the white-trained algorithm, we provide additional adjustments as needed.

Policy Implications

Insofar as conventional machine learning procedures do not produce the accuracy and fairness that some stakeholders require, it is possible to alter conventional practice to respond explicitly to many salient stakeholder claims even if they are unsupported by the facts. The results can be a politically acceptable risk assessment tools.

Keywords

risk Assessment, machine learning, forecasting, racial bias, fairness

Direct correspondence to Richard Berk, Department of Criminology, McNeil Hall, University of Pennsylvania, Philadelphia, PA. (e-mail berkr@sas.upenn.edu)

Introduction

Driven largely by computer scientists, statisticians, and legal scholars, the literature on fairness for algorithmic, criminal justice risk assessments is large and rapidly

growing (e.g., Star, 2014; Goel et al., 2016; Friedler et al., 2016; Chouldechova, 2016; Doleac and Stevenson, 2106; Hamilton, 2016; Kleinberg et al., 2017b; Corbett-Davies et al., 2017; Berk et al., 2018; Goal et al., 2018; Hug, 2019; Mayson, 2019). Many of the issues are complex. In particular, there are provably, inherent tradeoffs between different kinds of fairness and between fairness and accuracy (Kleinberg et al., 2017a, Chouldechova, 2017).[1] Despite well-intended aspirations, you can't have it all.

Proposed technical solutions typically select one or two kinds of fairness for which a "fair" algorithm can provided. Other forms of fairness and the fairness tradeoffs are ignored (Corbett-Davies, 2018). Reductions in accuracy are commonly an afterthought. There is, moreover, no single, dominant kind of fairness. Different stakeholders stubbornly can hold different and legitimate conceptions of fairness. Too often, gridlock is the result.

No clear resolution is likely in the near term. Meanwhile, criminal justice decisions will be made for many thousands of offenders. Various forms of risk assessments commonly will inform those decisions. In this paper, we propose an algorithmic fallback position that might be applied immediately to the construction of risk assessment tools. Rather than fair risk assessment, we offer, as a demonstration of concept, *politically acceptable* risk assessment. Perhaps a politically acceptable risk assessment approach can break the gridlock.

[1]Berk (2018a) provides a short and very accessible primer.

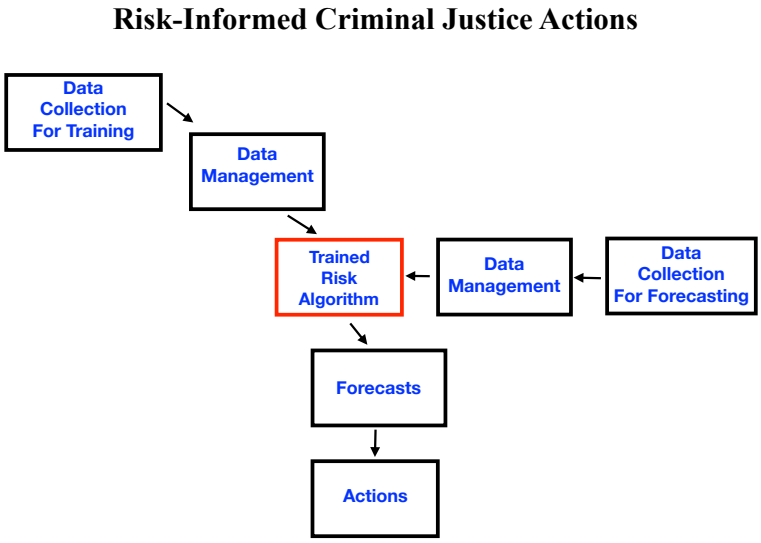

Figure 1: Risk Assessment components From Data Collection to Actions Taken

1 A Broader View of Risk Assessment

Much of the controversy over risk assessment conflates several related processes that make informed discussion extremely difficult. In particular, an "algorithm" often is blamed when by itself it may introduce no unfairness whatsoever. Both critics and supporters sometimes fail to appreciate that risk algorithms sit within a larger set of pursuits, any of which may be a source of unfairness.

Figure 1 provides an overview of the training and use of criminal justice, risk algorithms. The process begins with the collection of data, and the management of those data, used to train the algorithm. For example, an arrest may be recorded on an electronic "rap sheet" that, in turn, is stored under an offender's unique identification number. Such data may be fully accurate and properly curated. Alternatively,

the data may include information that some stakeholders label as "biased." Stop-and-frisk policing, for instance, is commonly blamed for inflated arrests counts for individuals from disadvantaged neighborhoods even when the scientific evidence can be at best equivocal (Grogger and Ridgeway, 2006, Ridgeway, 2006). Yet, racial animus can play a role, often at the level of individual police officers (Ridgeway and MacDonald, 2014). In short, criminal justice data sometimes can be a root cause of risk assessment unfairness, but the issues sometimes are subtle.

Modern, algorithmic risk tools are "trained" in the sense that they inductively seek associations between predictors (e.g., prior record) and an outcome of interest (e.g., a re-arrest). There is no model.[2] The associations are used to construct a measure of risk. Risk can be represented as a numerical score, a probability of one or more untoward outcomes, or a particular outcome class. Because criminal justice decisions are typically categorical (e.g., release on parole or not), often built into the algorithmic is machinery to translate a numerical score or probability into an appropriate categorical outcome. If not, some less formal means are required to generate an outcome class to serve as the forecast. A probability of a re-arrest, for instance, is not actionable until it is high enough to warrant assignment to the outcome class of, say, "recidivist."

Once training is completed, the risk algorithm can be used to forecast outcomes for new cases when their outcomes are unknown. Data for such forecasting must be collected and properly managed, just as done for the training data. The process can

[2]A model is an algebraic theory of how the world works. An algorithm is a computational procedure. Force = Mass × Acceleration is a model. When you balance your checkbook, you are applying an algorithm. The two should not be confused. They can have very different properties. Breiman (2001a) provides a telling discussion that is worth a careful read.

be demanding because the data used for forecasting must have the same predictors as the data used for training and be realized in the same fashion. Instructive forecasts might well be obtained, for example, from training data and forecasting data properly realized from the same jurisdiction, in roughly the same time period, and for the same criminal justice setting (e.g., parole decisions). And just as for the training data, there will usually be important data quality concerns.

The forecasting data are provided to the trained algorithm that, in turn, produces forecasts for each case. Ideally, these are forecasted outcome classes such as a post-release arrest for a violent crime, a post-release arrest for a non-violent crime, or no post-release arrest whatsoever. Often, the reliability of those forecasts also can be determined (Berk, 2018b). With the forecasts and (preferably) reliabilities in hand, decisions can be made and actions taken.

The recent risk literature makes clear that machine learning, algorithmic methods can have demonstrable, superior performance compared to subjective or model-based methods (Berk and Bleich, 2014; Berk et al., 2016). It has been long known that even very simple statistical methods produce better accuracy than subjective approaches (Meehl, 1954; Dawes et al., 1989), and as a mathematical matter, machine learning algorithms can adaptively find far more complex relationships in the data, should they exist, than conventional models such as logistic regression (Hastie et al., 2009). When the relationships between predictors and an outcome are not complex, algorithmic methods will perform no worse than conventional models.

To the best of our knowledge, no one has shown that risk assessment bias in criminal justice settings is produced by algorithmic code itself, but biases already

present in the data can be carried foreword by that code. In response, there is a substantial literature on how the certain data-generated biases can be moderated and even in some circumstances removed (e.g., Kamiran and Calders, 2009; Kamishima et al., 2011; Kamiran et all., 2012; Feldman et al., 2015; Hardt et al., 2016 ; Berk et al., 2017; Kearns et al., 2018a,b). In practice, it has proved impossible to remove all bias in part because of the inevitable tradeoffs mentioned earlier and discussed in much more detail below.[3]

Even significant biases are harmless by themselves. As Figure 1 is meant to emphasize, harm can only result when those biases influence actions. Wrongfully detaining an offender at an arraignment can be one instance.[4]

Efforts to reduce and perhaps even eliminate real harm can be undertaken at any of the stages shown in Figure 1. In this paper, we consider all of the steps except actions. Substantial success in any of these pursuits can be passed along to subsequent steps with the understanding that, nevertheless, *other* sources of biased may be introduced at those points. There have been proposals to address some of these other sources (Ridgeway, 2016), but as important as these are, they are downstream from the potential remedies considered below.

[3]All of these definitions of fairness address inequalities between groups. There is also a literature in computer science and statistics on inequalities between individuals. The basic idea is that similarly situated individuals should be treated similarly (Speicher et al., 2018).

[4]Some argue that because at an arraignment there has yet to be a conviction, the term "offender" is misleading. "Arrestee" or "suspect" is perhaps more accurate, but offender is the term commonly used. The term "defendant" is inappropriate because many of those arraigned will not officially charged with a crime beyond what is reported by the arresting officers.

Politically Acceptable Risk Assessment

The major forms of unfairness are addressed in some detail shortly. As already emphasized, there are several different kinds that can be incompatible and about which there can be strongly held views. In the mix also are well-founded concerns about the role of race in the American criminal justice system. Just as for fairness itself, the issues are complicated. Apprehensions by the criminal justice system are often the last stop on a train that left the station early in an offender's childhood. "Mass incarceration," for example, is a product of many factors, including widespread exposure to violence early in life (Bloom, 2014; Heissel, 2017) and many other kinds of disadvantage in which race-related processes are implicated. These too can produce strongly held views.

Properly understood, risk algorithms are meant to help inform criminal justice decisions to make them more accurate and more fair, but they cannot be expected to reverse generations of disadvantage. Algorithms will surely make some forecasting errors and will surely produce some unfairness, but the aspiration is improved accuracy and fairness *compared to current practice.* Initial evidence indicates that both can be achieved in real settings (e.g., Berk et al., 2016, Jung et al., (2017); Kleinberg et al., 2017b; Berk, 2018b; Goel et al., 2018; Guay and Parent, 2018; Mullainathan, 2019).

Nevertheless, in the fraught larger context, risk algorithms can elicit a primal response too often supported by deep misunderstandings. Figure 2 illustrates one instance. The existing risk assessment tool had in fact been a great success, and justified far less intrusive probation supervision for low risk offenders, who were dis-

Figure 2: Announcement For An Anti-Risk Assessment Rally

proportionately African-American (Berk et al., 2010). Moreover, nearly a decade earlier, criminal justice officials had decided on their own not to use zip code as a predictor. They were concerned about even the appearance of unfairness. This decision had been explained several times to social justice advocates in public settings, but that account was by and large ignored.

Given deep and often unbending opposition from some stakeholders and the impossibility of a perfectly accurate and absolutely fair risk instrument, it might make sense to focus instead on algorithmic risk instruments that are *politically acceptable*. To simplify the exposition, we will proceed assuming offenders are either African-American or White. Similar issues can arise for other racial and ethnic groups, and our approach easily generalizes.

Claims often are made that criminal justice decisions are tainted by "white privilege," among other racial factors (Harcourt, 2007; Star, 2014; Tonrey, 2014; Goel et al., 2016; Ferguson, 2017). The issues are actually a philosophical swamp (Sowell, 2002; Boonin, 2011), but most stakeholder arguments, based on social constructions, are not intended to be subtle; the enterprise primarily is political. One might infer, therefore, that if Black offenders were sentenced in the same manner as White offenders, Black offenders would demonstrably benefit, and a form of equality would result. Moreover, no Whites would be harmed because none would be made worse off.[5]

A practical response in the algorithmic world would be to train a risk algorithm solely on White offenders and then compute from test data the risks for Black and White offenders separately using the white-trained instrument. Black and Whites jointly would benefit from white privilege. We start there after laying a bit of additional conceptual and empirical groundwork.[6]

A key implication is that a politically acceptable risk assessment is simply *a risk procedure on which a sufficient number of stakeholders can agree.* In some cases, this can be through deliberations of a city council or other legislative bodies. In other cases, reform efforts are guided by a standing committee of stakeholder representatives, a commission, or an official advisory board. In yet other cases, there can be

[5]This is rather different from college affirmative action efforts in which some Whites, who otherwise would have been admitted, are made worse off.

[6]To anticipate, a machine learning algorithm such as random forests or stochastic gradient boosting could be used to fit data only from white offenders. With the trained algorithmic results in hand, risk forecasts would be computed separately using test data for White offenders and Black offenders. This approach is related to procedures used to partition wage differentials by gender (Binder, 1973; Oaxaca, 1973), which was invented in demography about 20 years earlier (Kitagowa, 1955).

ad hoc oversight committee with wide stakeholder representation. In practice, there also are a host of details to be worked out such as which stakeholders can participate and the procedural rules to be adopted. Further discussion would be a lengthy diversion and beyond the expertise of the authors. Accessible treatments are easily found (e.g., Dovovan et al., 2014).

"Sufficient number " will vary over settings, organizational structures, and issues, but is manifested in a decision about whether or not to proceed with a particular risk assessment method, even if there is some residual dissatisfaction, grumbling and even opposition. Politically acceptable risk assessment has no deeper meaning. There are no necessary links to the far more weighty thinking about justice in the works of scholars such as John Rawls (2001) or Amartya Sen (2009).[7]

Data To be Analyzed

We analyze a dataset of 300,000 offenders from a large metropolitan area arraigned very soon after an arrest from 2007 through June of 2015. One can certainly wonder whether too many years are included. There is no reason to assume that the administration of arraignments, the governing statutes and mix of offenders had been constant over that time period. But, arraignments over those years show remarkable inertia and in other, ongoing work, we have found little change in the structuring of

[7]There are no quick and easy answers here either and certainly no conceptual consensus. There is even work by Nobel Laureate Kenneth Arrow formally proving that when voters have more than two policy options and ranked preferences between them, this is no ranked voting system that can translate individual policies preferences into an appropriate collective result meeting several sensible conditions, such as the independence of irrelevant alternatives and preference transitivity (Arrow, 1951).

risk for subsets of years. For issues addressed in this paper, working with the full dataset seems to pose no important problems and the very large number of observations is extremely helpful; if one is looking for subtle manifestations of racial bias, "big data" is beneficial.

For the full dataset, 67% of the offenders are African-American, and 32% of the offenders are White. We use an outcome variable with three classes: no arrest after an arraignment release, an arrest for a non-violent crime after an arraignment release, and an arrest for a violent crime after an arraignment release. These releases assume a return for a subsequent court appearance and are the three outcome classes for which risk will be computed.

After release, 57% of the offenders had no arrest, 33% had an arrest for a non-violent crime, and 10% had an arrest for a violent crime. The corresponding figures for Whites alone are 58%, 35%, and 7%. For Blacks, the corresponding figures are 56%, 32% and 11%. These figures define, respectively, the base rate for the full dataset, the base rate for White offenders, and the base rate for Black offenders. Figure 3 shows the base rate distributions for Whites and Blacks, which to the eye look rather similar. We will see later that base rates play a key role in the determination of fairness.

Perhaps the most notable difference by race is the likelihood of arrests for a violent crimes. Some might see this as an important difference in base rates between White and Black offenders. Because the percentages are relatively small, the disparity sensibly could be represented as a ratio: $.11/.07 = 1.57$. Black offenders are greater

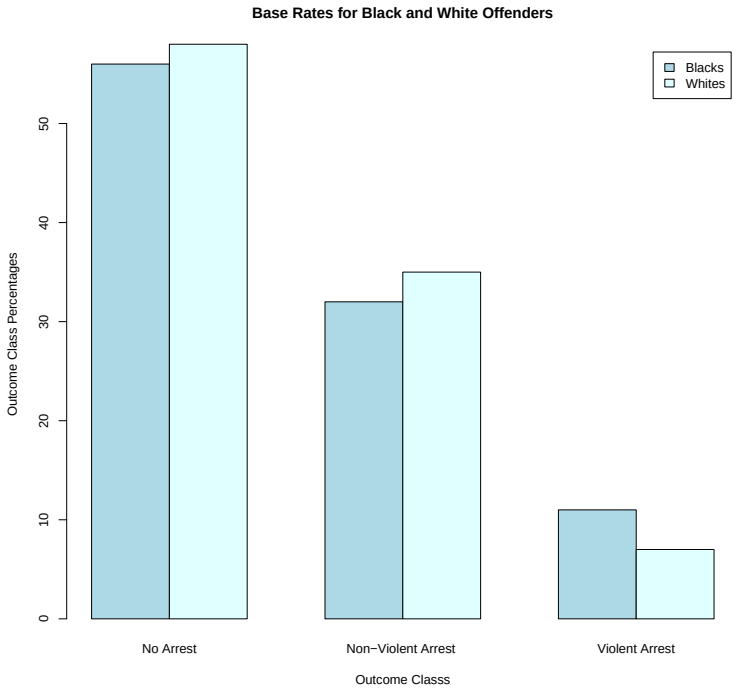

Figure 3: Post-Release Arrests Base Rates for Black and White Offenders with Proportions on The Vertical Axis

than 50% more likely to be arrested for a violence crime.[8]

Such racial comparisons also can imply differential accuracy in *current practice*. Presumably, a magistrate would be very reluctant to release an offender who would be re-arrested for a violent crime. Yet, that apparently is what happens 7% of the time for White offenders and 11% of the time for Black offenders. More such mistakes are made for Black offenders, and because violent crimes typically are intra-racial, these mistakes disproportionately affect Black victims of violent crime.[9]

Table 1 shows the predictors used and the usual summary statistics. These predictors were routinely available in machine readable form, which means they would be available to train the algorithm and later to be used at arraignments when forecasts were needed. Age is only biographical variable included as a predictor. Race, ethnicity, gender, education and marital status were available in the data provided but were not used to train the risk tool out of an abundance of caution about fairness, as articulated by stakeholders. These decisions may be re-visited.

A variety of priors were available from rap sheet that represent counts of earlier arrests.[10] All had long right tails (i.e., compare the median to the mean) and on

[8]The adjectives"non-violent" and "violent" quite accurately describe the vast majority the crimes in their respective categories. But there are a few kinds of crime in the violent crime category that some criminal justice stakeholders wanted to include because of heightened public concerns, even if not violent in themselves. The most common examples were "sex crimes" such as child pornography, which exploit children for sexual stimulation. No assault need be involved. This decision may be re-visited in future discussions among stakeholders.

[9]Because of the very large number of offenders arraigned in a given year, the difference between 7% and 11% in current practice translates into several thousand more victims of violent crime who will be disproportionately African-American.

[10]Some argue that convictions are a better measure of criminal activity. However, roughly 90% of all cases are resolved through a plea bargain, and the negotiated plea often has little correspondence to actual crime that may have been committed. Moreover, there is no reason to assume that race plays no role in the bargaining process. And as an empirical matter, we have found that arrests are stronger predictors than convictions when the intent is to forecast arrests.

Predictor	Min	Q1	Q2	Mean	Q3	Max
Age	15.01	23.15	30.24	32.85	41.08	87.89
Murder Priors	0.00	0.00	0.00	0.127	0.00	44.00
DV Priors	0.00	0.00	0.00	0.107	0.00	9.00
Prior Sentences	0.00	0.00	0.00	1.816	2.00	121.00
Property Priors	0.00	0.00	0.00	7.103	7.00	1137.00
Serious Priors	0.00	0.00	0.00	2.354	3.00	275.00
Violence Priors	0.00	0.00	0.00	6.372	8.00	449.00
Sexual Priors	0.00	0.00	0.00	0.279	0.00	275.00
Firearm Priors	0.00	0.00	0.00	1.423	0.00	215.00
Weapons Priors	0.00	0.00	0.00	2.039	2.00	215.00
First Charge Age	12.01	18.77	20.87	24.40	26.91	87.89
Any Charge Count	1.00	2.00	3.00	4.604	5.00	40.00
Murder Count	0.00	0.00	0.00	0.029	0.00	4.00
Weapon Count	0.00	0.00	0.00	1.071	1.00	33.00
Property Count	0.00	0.00	0.00	1.104	1.00	17.00
Drug Distribution Count	0.00	0.00	0.00	0.209	0.00	31.00
Domestic Violence Count	0.00	0.00	0.00	0.433	0.00	9.00
Violence Count	0.00	0.00	0.00	1.514	2.00	36.00
Serious Count	0.00	0.00	0.00	0.486	0.00	15.00
Sexual Count	0.00	0.00	0.00	0.113	0.00	63.00
Firearm Count	0.00	0.00	0.00	0.294	0.00	13.00

Table 1: The Minimum Value, First Quartile, Median, Mean Third Quartile and Maximum Value for Each Predictor

occasion, outliers that criminal justice officials claimed were not recording errors. From the priors, we were able to compute that earliest age at which an offender was charged as an adult. Machine learning risk tools have established that offenders who start early are generally higher risk (Berk, 2018b). The remaining predictors are the current criminal charges alleged by arresting police officers prior to arraignment and typically acceptable to participating prosecutors. Here too, the predictors had long right tails with occasional outliers that were alleged to not be errors.[11]

The predictors available at an arraignment are generally very limited and often substantially correlated with one another. For example, serious priors and violence priors significantly overlapped. The first is essentially all major felonies while the second largely includes crimes of violence only.[12]

We anticipated concerns about predictors derived from arrests recorded on rap sheets. One common complaint is that Black offenders often have longer rap sheets because of police "bias." Those longer records, in turn, lead to unfavorable risk assessments. Police racial animus is commonly cited to explain arrests that are unjustified. A related claim is that Black neighborhoods are "over-policed," even if the over-policing is well-intended; with more police, there will be more arrests.

The issues are complicated and to date are not empirically resolved. Moreover, general conclusions are essentially beside the point because in practice, data quality

[11] A few outliers are usually not a serious problem for non-parametric procedures of the form used below in part because they are thoroughly dominated by the very large number of observations that not outliers and in part because the functions fitted are generally robust to outliers. This will be apparent later.

[12] Because interested centers on forecasting, not explanation, substantial dependence between predictors is usually not an important concern. There is little interest in the the usual covariance adjusted predictors.

must be addressed locally and for particular criminal justice measures. We take, therefore, no position on the relevant facts in this paper. For example, there are no doubt some police officers who unfairly target individuals in predominantly Black neighborhoods. The question is whether such practices are frequent enough to impact arrest statistics in a significant manner. With respect to over-policing in major metropolitan areas, police deployments are dominated by calls for service (911 calls), and citizens in disadvantaged neighborhoods disproportionately call the police. A higher density of police can be a consequence of these calls. One would be hard pressed to see that higher density as police exercising racial animus, although longer rap sheets for Black offenders could result.

Nevertheless, we discarded all predictors derived from prior arrests for less serious crimes even though such priors were included in the dataset. These are crimes for which police might exercise substantial discretion (e.g., loitering). Such discretion is said by some to introduce unjustified racial disparities and is an easy target for risk assessment critics. We also excluded all priors for arrests as a juvenile, which were also available. The reasoning was similar but, in addition, responded to claims that crimes committed by juveniles were less about criminal proclivities and more about poor self-control and bad decisions. We retained priors for arrests as an adult that were likely to be charged as major felonies, often associated with violence and identifiable victims. In the interest of comparability, all of predictor exclusions and retentions were made for Black offenders and White offenders alike.

Training data and test data were constructed as disjoint random splits of equal size. We planned to use stochastic gradient boosting, which is vulnerable to over-

fitting. Therefore, valid test data were essential.[13] In addition, we anticipated subsetting the data by race. Implementing both kinds of data partitioning required beginning with a large number of observations. Because the full dataset included 300,000 observations, each of the analyses to follow was undertaken with at least several thousand cases.

Statistical Methods Employed

Stochastic gradient boosting (Friedman, 2001; 2003; Hastie et al., 2009 Section 10.10), implemented in R as *XGBoost*, was used for the analysis.[14] The target cost ratios were to be the same for all classification errors. By properly weighting the data before the analysis began, all empirical cost ratios were, as intended, approximately 1 to 1.[15] The 1 to 1 target cost ratios were a provisional decision made by criminal justice officials responsible for arraignments, and the cost ratios to be used

[13]Machine learning methods are adaptive and subject to significant overfitting. Recommended practice counsels using one dataset (i.e., training data) for learning and another dataset (i.e. test data) to obtain "honest" output. The two datasets should be IID realizations from the same joint probability distribution or a very good proxies thereof.

[14]Gradient boosting is a powerful machine learning procedure. A very large number of passes are made sequentially through the training data. With each pass, a decision tree is grown. Observations are re-weighted so that cases with larger residuals from the immediately preceding decision tree are given more weight; the algorithm then works harder to fit cases that previously were fit less well. The fitted values from each pass ultimately are combined in a weighted, linear fashion, which for categorial response variables outputs fitted probabilities of risk for each outcome class. Ridgeway (2007) provides a useful overview although he does not discuss the multinomial formulation we use here.

[15]In the binary case, the cost ratio is the number of false positives divided by the number of false negatives, or the inverse. When there are more than two outcome classes, there are no conventional naming conventions, but the cost ratio is still the ratio of the number of classification errors for a pair of different outcome classes. Cost ratios shows how the algorithm is trading off one kind of classification error against another. More details are provided shortly.

in practice are still to be determined.[16] Arriving at appropriate cost ratios is a very important component of any classification enterprise, but they are peripheral to the issues addressed in this paper.[17]

Confusion Tables and Definitions of Fairness

Just as in common practice, the results of our machine learning, training exercises with categorical outcomes will be shown in confusion tables. Such tables are nothing more than a cross-tabulation of the actual outcome classes and the fitted outcome classes, ideally constructed from test data to counter overfitting. Important for our purposes, these tables define the most common forms of unfairness, although there are no universal naming conventions for them.

Table 2 provides a vehicle for the discussion. There are just two outcomes: success and failure. Confusion tables can be constructed in much the same fashion when there are more than two outcome classes, but that would unnecessarily complicate the discussion. The letter in each cell represents the number of cases. For example, the letter "a" is the number of cases for which the fitted class is a failure, and the

[16]Target cost ratios are specified *a priori* as a policy matter. The gradient boosting procedure is tuned during training to approximate a target cost ratio. The approximation achieved is the empirical cost ratio shown in a confusion table.

[17]The technical literature on risk assessment is evolving so quickly that commentaries are often badly dated soon after they are written. A heavy burden for critical reading falls on consumers of such commentaries. For example, the recent report "Algorithms in the Justice System: Some statistical issues," written for the Royal Statistical Society in 2018, implicitly discards some of the most effective and mathematically justified machine learning tools available when it asserts without any justification that risk should be in probability units. Neither random forests (Breiman, 2001b) nor support vector machines (Boser et al., 1992) output probabilities. More immediately relevant for this paper, although the fitted classes obtained from stochastic gradient boosting often forecast well, the fitted probabilities are pushed toward 0.0 or 1.0, making them terrible estimates. In effect, the probability estimates are useless as ends in themselves (Mease et al., 2007).

actual class is a failure. These cases that are correctly classified as failures. The letter "b" is the number of cases for which the fitted class is a success, and the actual class is a failure. These cases are incorrectly classified as successes. The letters "c" and "d" have analogous interpretations.

	Fitted Failure	Fitted Success	Classification Error
Observed Failure	a	b	$b/(a+b)$
Observed Success	c	d	$c/(c+d)$
Forecasting Error	$c/(a+c)$	$b/(b+d)$	Overall Error $= \frac{(b+c)}{(a+b+c+d)}$

Table 2: A Stylized Confusion Table for a Binary Outcome Variable. (The letters represent counts of the number of observations. There are two outcomes: "success" and "failure.")

The lower right cell contains the proportion of cases for which the fitted outcome class and the observed class is not the same. This often is called overall classification error. It treats the costs of all such errors the same; the cost of incorrectly classifying a failure as a success is the same the cost as incorrectly classifying a success as a failure. Usually the costs differ, sometimes dramatically. A bit more discussion follows shortly.[18]

There is typically far more interest in the calculations along the margins of the table, whose calculations shown in Table 2 are unaffected by an assumption of equal costs. Nevertheless, if the overall classification error differs for, say, White offenders and Black offenders, the algorithmic results can be criticized as unfair. More classification errors are made for one group compared to the other.

[18]These are not necessarily monetary costs and an extended discussion is for this paper a diversion. A very accessible treatment can be found in Berk's monograph (2018b).

Under "Classification Error" are the conditional proportions for the actual failures and actual successes, respectively, whose fitted class is not the same as the observed class. Given an observed failure, what fraction of the time does the classification procedure claim a success? Given an observed success, what fraction of the time does the classification procedure claim a failure? If a failure is called a "positive" and a success is called a "negative," $b/(a+b)$ is the false negative rate and $c/(c+d)$ is called the false positive rate, which can be expressed as a proportion or a probability depending on how the data were generated.[19]

The designation of a success as negative and a failure as positive is perhaps counter-intuitive, but often the primarily intent of a risk assessment is to help prevent failures (e.g., a re-arrest for a violent crime). A prerequisite is to correctly forecast failures, which is beneficial. Hence, it is a positive. The reasoning for designating a success as a negative is analogous.

The false positive and false negative rates are commonly a major focus when concerns about fairness are raised. Suppose successes and failures refer to the absence of a post-release arrest or the presence of a post-release arrest respectively. If the false negative rate is higher for White offenders than Black offenders, say, the risk assessment is more likely to incorrectly classify White offenders as good risks than Black offenders. If the false positive rate is higher for Black offenders compared to White offenders, say, the risk assessment is more likely to incorrectly classify Black offenders as bad risks compared to White offenders. In both instances, many would

[19]To justify using proportions as probabilities, researchers common try to make the case that the data were realized IID from some relevant joint probability distribution. For an in depth discussion, see Berk (2018b).

argue that classification errors unfairly favor White offenders. Some would call this *"inequality of opportunity."* Even before an algorithm is used to help inform criminal justice decisions, classification accuracy is better for one group than another.

The false negative and false positive rate assume that the true outcome class is known. When risk tools are actually used to make forecasts, the true outcome class is not known – that's why forecasts are needed. Therefore, interest naturally turns to forecasting errors. Under "Forecasting Error," one conditions on the assigned class, not the truth. Thus, one has measures of forecasting error $c/(a+c)$ and $b/(b+d)$. Given a fitted class of failure, what fraction of the time is the forecast incorrect? Given a fitted class of success, what fraction of the time is the forecast incorrect?

If the table is assembled from proper test data, both fractions properly can be interpreted as the proportion of times the *forecasted* class is incorrect. These error rates help inform decision-makers about how the forecasting procedure will perform in practice. How reliable is a forecast of success? How reliable is a forecast of failure? Forecasting error arguably is the most important performance criterion in criminal justice risk assessments, and figures significantly in discussions of fairness. When a forecast of failure is made, is that forecast equally accurate for Black offenders and White offenders? When a forecast of success is made, is that forecast equally accurate for Black offenders and White offenders? For example, if a forecast of failure is less accurate for Black offenders, Black offenders will experience a greater number of incorrectly forecasted post-release arrests; they will be more likely to be labeled in error as high risk. Some would call this an absence of *"predictive parity."*

The quotient of b/c is called the cost ratio, and as noted earlier, captures how the

algorithm is trading false negatives against false positives. For example, if $b = 500$ and $c = 250$, the cost ratio of false negatives to false positives is 2 to 1. This means that one false positive is "worth" two false negatives; a false positive is treated as having twice the cost of a false negative. The algorithm then works harder to reduce false positives than to reduce false negatives. Ideally, cost ratios should be the same for Black offenders and White offenders because the algorithm then is making the same tradeoffs for both. An absence such equality some would call *"differential treatment."* But to date, cost ratios have received little interest in the fairness literature.

Finally, and often the first place stakeholder look to identify unfairness, is that fraction of cases forecasted to fail and the faction of cases forecasted to succeed: $(a+c)/(a+b+c+d)$ and $(b+d)/(a+b+c+d)$ respectively. A common complaint is that a larger percentage of Black offenders are forecasted to fail compared to White offenders. Some would call this *"inequality of outcome."* But what if one group is actually more likely to fail than another? That is, what if the base rates truly differ? Should that not be represented in the forecasting outcomes? We will return to this important issue later.

Considering in any depth the necessary tradeoffs between different kinds of fairness and between fairness and accuracy would require a far more technical discussion that is peripheral to goals of this paper and can, in any case, be found in other sources (e.g., Berk et al., 2018). Two brief points may suffice.

First, machine learning algorithms optimize some loss function such as the deviance. Any constraints placed on the fitting process will increase the loss and reduce

forecasting accuracy; optimality is forfeited. There is a necessary tradeoff between forecasting accuracy and each kind of fairness.

Second, inevitable tradeoffs between different kind of fairness have been proved (Kleinberg et al., 2017a, Chouldechova, 2017). In particular, if "the base rates differ across groups, any [risk] instrument that satisfies predictive parity at a given threshold ... *must* have imbalanced false positive or false negative errors rates at that threshold" (Chouldechova, 2017: 5 – emphasis in the original).[20]

There can be stylized exceptions. One example is a perfect fit to the data; one has absolutely not classification errors (b and $c = 0$). In practice, however, the message is that if the base rates differ between groups, one cannot have the same forecasting accuracy for each group *and* the same false positive and negative rates for each group. There necessarily will be an absence of predictive parity or an absence of equal opportunity.

A key implication for this paper is that base rates really matter. The pervasiveness of different base rates for different groups is an invitation to widespread unfairness. With different base rates, there effectively can be no technical solution to bias in which each kind of unfairness disappears. Compromises are needed, which in the current political climate presents daunting challenges. We aim to help.

[20] "Threshold" refers to the probability of failure (or success) used to assign an outcome class, such as probabilities of a failure (or success) greater than .50.

Results

The analyses to follow use stochastic gradient boosting with different forms of the data to address some ways that one might moderate various kinds of unfairness that have been considered in the risk assessment literature. In each analysis, we will apply the boosting algorithm tuned in the same fashion and compare the confusion tables for White and Black offenders. The more similar the confusion tables, the less evidence there is for bias.

A Conventional Risk Analysis Addressing Fairness

To set the context, we first applied stochastic gradient boosting to *the entire training dataset* and then constructed confusion tables from test data separately for White offenders and Black offenders.[21]

Table 3 is the confusion table for Whites. 54% of the White offenders were predicted not to be re-arrested after an arraignment release. 33% were predicted to be re-arrested for a non-violent crime. 12% were predicted to be re-arrested for a violent crime.[22]

Classification error is not especially relevant in this policy setting. As already

[21]Consider a conventional linear regression. The model's parameter values are estimated using all of the training data. Then predicted values are obtained separately for White and Black offenders by simply inserting test data for Whites back into the linear model results followed separately by inserting the test data for Blacks back into the linear model results. More formally, recall that $\hat{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. With $\hat{\boldsymbol{\beta}}$ *already estimated for the full training dataset and then fixed*, one can separately insert from test data $\boldsymbol{\mathcal{X}}_{Whites}$ and $\boldsymbol{\mathcal{X}}_{Blacks}$, instead of training data $\mathbf{X}$, to get the fitted values for White and Black offenders respectively. Similar steps are involved here. Symbolically, just substitute $\hat{f}(\mathbf{X})$ for $\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

[22]Recall that all of offenders were released with the expectation that they would be returning for a later court appearance. Charges were not dropped at the arraignment.

Table 3: Stochastic Gradient Boosting Confusion Table from Test Data for White Offenders Using the Conventionally Trained Algorithm: 54% Predicted No Arrest, 33% Predicted Non-violent Arrest, 12% Predicted Violent Arrest

Observed Outcomes	No Arrest Predicted	Non-Violent Arrest Predicted	Violent Arrest Predicted	Classification Error
No Arrest	17877	6848	2535	.34
Non-Violent Arrest	6454	7593	2062	.53
Violent Arrest	1859	1779	1234	.75
Prediction Error	.32	.53	.79	

noted, it assumes that each outcome is known and then computes the fraction of cases that the algorithm misclassifies. When a decision needs to be made about a particular offender, the outcome is not known. On cannot tell in which row of the confusion table the offender belongs.[23] When the observed outcome is no arrest, that outcome is misclassified 34% of the time. When the observed outcome is an arrest for a non-violent crime, that outcome is misclassified 53% of the time. When the observed outcome is an arrest for a violent crime, that outcome is misclassified 75% of the time.[24] But which applies to any given offender when a decision is to release or detain?

The differences between the three misclassification proportions is a routine disparity that can result from an unbalanced base rate distribution. It is usually more difficult to correctly classify outcomes that are substantially less common. If desired,

[23]When there are more than two outcome classes, the terms "false positive" and "false negative" make no sense. For each actual outcome class, there can be two or more ways to misclassify. And with more than two classes, if one is called a "positive" and one is called a "negative," what are the other classes to be called? As far as we know, there are no naming conventions that have addressed this setting.

[24]Higher misclassification error is common for relatively rare outcome classes. How this can be properly handed in practice depends on the setting (e.g., Berk et al., 2016)

one can weight the less common cases relatively more heavily, but that option was precluded here because the provisional 1 to 1 cost ratios would not longer hold.

Far more important for policy is forecasting accuracy when the outcome is *not* known. With training and test data, generalization error can be estimated.[25] From Table 3, a forecast of no arrest is wrong 32% of the time. A forecast of an arrest a non-violent crime is wrong 53% of the time. A forecast of an arrest for a violent crime is wrong 79% of the time. With different cost ratios, it would be possible to do somewhat better, but there is still a demonstrable improvement in predictive accuracy. Applying a Bayes classifier to the marginal distribution of the outcome, one would forecast no arrest using none of the predictors and be wrong 43% of the time. If no arrest is forecasted from Table 3, the estimate of generalization error drops to 32%. That is a reduction in generalization error of about a quarter compared to current practice. This substantial improvement in predictive accuracy is achieved despite the paucity of predictors available and potentially powerful predictors, such as gender, which were discarded.[26]

Table 4 is the confusion table for Black offenders constructed from test data. 55% of the Black offenders were predicted to not be re-arrested after an arraignment release. 33% were predicted to be re-arrested for a non-violent crime. 12% were predicted to be re-arrested for a violent crime. The figures are almost identical to as those for White offenders. Black offenders are predicted to be no more or no less

[25]For categorical outcomes, the definition of generalization error depends on whether the fitted value is a class or a class probability (Hastie et al., 2009: 221). We are applying the definition for fitted classes. For a 0-1 loss, the usual MSE becomes the proportion forecasted incorrectly.

[26]Given the large number of offenders arraigned in a given year, this could translate into thousands fewer crime victims.

Table 4: Stochastic Gradient Boosting Confusion Table from Test Data for Black Offenders Using the Conventionally Trained Algorithm: 55% Predicted No Arrest, 33% Predicted Non-violent Arrest, 12% Predicted Violent Arrest

Observed Outcomes	No Arrest Predicted	Non-Violent Arrest Predicted	Violent Arrest Predicted	Classification Error
No Arrest	38178	14301	5098	.34
Non-Violent Arrest	13346	15749	4231	.54
Violent Arrest	3792	3965	2817	.74
Prediction Error	.31	.54	.77	

risky than White offenders. There is virtually no support for claims of inequality of outcomes. The figures for classification error and prediction error are also nearly the same for Black offenders and White offenders. Hence, there is virtually no support for claims of inequality of opportunity or predictive inequality. One plausible explanation, discussed in more depth later, is that in this instance the base rates for Black offenders and White offenders are sufficiently alike. Because all offenders had been very recently arrested, they may be more homogeneous than samples of all Black and White individuals from the jurisdiction.

But some caution is necessary. The near equivalences should not be taken too literally. Even if one treats these very large samples as populations, there is some imprecision in all of the proportions computed from the confusion tables. The main source is that tuning machine learning algorithms is an approximation enterprise. Small differences in tuning parameter values can sometimes translate into variation of several percentage points when confusion tables are constructed and proportions computed. In addition, even though the risk assessments for Blacks and Whites were undertaken with the exact same trained algorithm, different offenders bring different

predictor distributions to the test data, which can produce disparities between the two confusion tables depending on the way the algorithm is tuned. For example, a greater number of violent crime prior arrests for Blacks than Whites could affect the results. Nevertheless, it would be difficult to make a strong case for racial unfairness from Tables 3 and 4.

The comparisons between Table 3 and Table 4 illustrate three points. First, despite widespread claims of algorithmic bias, there is no manifest evidence of unfairness from this risk tool, at least for the machine learning method used, the provisional cost ratios imposed, the way the algorithm was tuned, and the conventional performance measures we computed. Racial bias in criminal justice risk assessments apparently is not inevitable.

Second, should there be important racial differences, even if an artifact of how the algorithm was trained, those differences will get ported into practice. The training data and the algorithmic structure resulting from the training typically are not revisited. It can be very important, therefore, to examine the robustness of the results through several rounds of retuning coupled with new random splits of the data into training and test subsets. When this was done for these data, some racial differences surfaced. Some favored Black offenders, some favored White offenders, but none of the racial disparities would likely be considered large in practical terms.

Third, results like those in Table 3 and Table 4 will not necessarily convince all stakeholders. Empirical evidence of equitable outcomes does not necessarily carry the day because of strongly held beliefs that the entire criminal justice system is rife with racial bias. How can such a system and the data it assembles produce anything

that is fair? Our empirical results could easily be dismissed as an aberration. From this perspective, the only reasonable action is to proactively reject all empirical risk assessments and rely instead on major structural and procedural reforms that achieve social justice. If stakeholders knew what kinds of structural and procedural reforms would secure full social justice, and if these stakeholders were confident that these reforms could be implemented quickly and with integrity, the proactive rejection of all risk assessment might have some merit. In short, one side sees the glass as half full and aims to add a bit more water. The other side sees the glass completely empty and seeks to start again with a new glass and a new water supply.

Training the Algorithm on White Offenders Only

Perhaps there is a constructive middle ground. One can take the claims of racial unfairness seriously and try to respond to them as risk assessment tools are built. In particular, one can frame claims of racial bias affecting Black offenders as implying that the treatment of White offenders is a manifestation of "White privilege." Algorithmic training can usefully respond to this view.

One starts by training the algorithm *only on White offenders*. Risk classification is then undertaken separately for Black and White offenders using their own test data. With the training only on Whites completed, the two sets of test data can be used for predicting different risk classes. We emphasize the algorithmic structure arrived at with the White training data is used separately with the White and Black test data. Risks for Black and White offenders are processed by the trained algorithm

as if everyone is White because that is all the algorithm "knows."[27]

Table 5: Stochastic Gradient Boosting Confusion Table from Test Data for White Offenders Using the White-Trained Algorithm: 59% Predicted No Arrest, 32% Predicted Non-violent Arrest, 9% Predicted Violent Arrest

Observed Outcomes	No Arrest Predicted	Non-Violent Arrest Predicted	Violent Arrest Predicted	Classification Error
No Arrest	24773	8030	2520	.30
Non-Violent Arrest	9063	9520	2203	.54
Violent Arrest	1841	1541	1033	.77
Prediction Error	.30	.51	.82	

Table 5 is the confusion table for White offenders. The results are very similar to those in Table 3. For example, 59% are predicted to not be re-arrested, 32% are predicted be re-arrested for a non-violent crime, and 9% are predicted to be re-arrested for a violent crime. This does not differ in important ways from the predicted risk classes when the algorithm was trained on all of the data. One might expect that accuracy would be better in Table 5 because the training data and test data are for White offenders. One possible implication is that the social and law enforcement practices responsible for recidivism are very similar for both racial groups. But we need to dig deeper.

Figure 4 shows for White offenders the fitting importance of each predictor. The prefix "A" denotes priors from arrests as an adult. The prefix "i" denotes charges from the instant crime for which an offender is being arraigned. Priors and counts are integers. Each predictor's contribution to the fit is computed for each pass through

[27]Training is undertaken with training data for Whites using $\hat{Y} = \hat{f}(\boldsymbol{X}_{Whites})$. Then with the function fixed, one can separately insert test data $\boldsymbol{\mathcal{X}}_{Whites}$ and $\boldsymbol{\mathcal{X}}_{Blacks}$ to get the fitted values for White and Black offenders respectively.

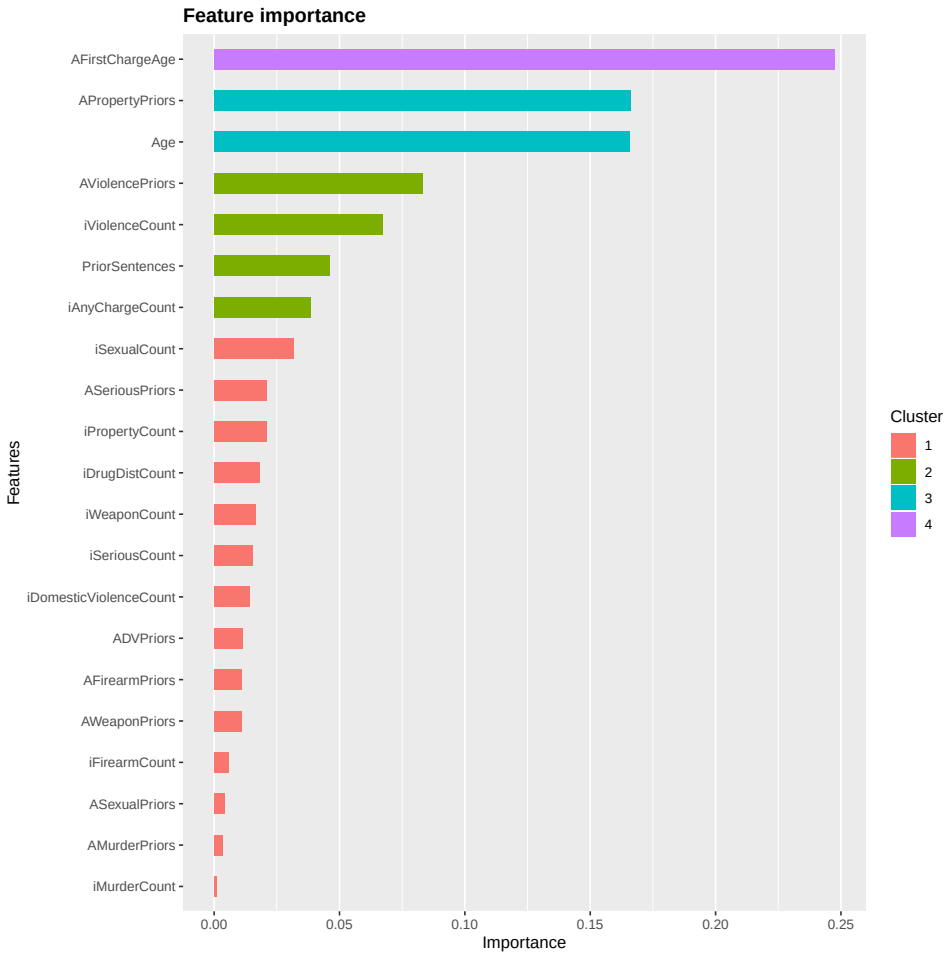

Figure 4: Input Contributions To Stochastic Gradient Boosting Fit for White Offenders Trained on White Offenders (contribution sum to 100%)

the data. Importance is computed as the average over passes, standardized so that the average contributions over all predictors sum to 100%. The formal rationale can be found in Hastie et al., (2009: section 15.3.2).

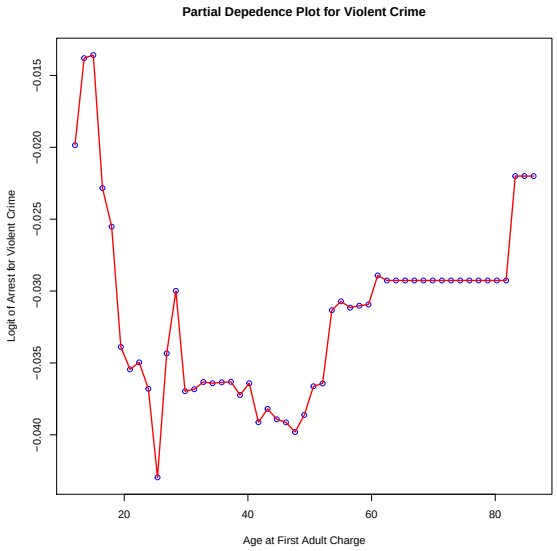

Figure 5: Partial Dependence Plot of the Association of Age at the First Adult Charge with An Arrest for a Violent Crime for White Offenders Using A White-Trained Trained Algorithm

The following three predictors, in order, dominate the fitting process. They will have important implications later.

1. **The age at which an offender is first charged as an adult** – For a juvenile to be charged as an adult requires that the crime responsible be very serious and typically, violent as well. Not surprisingly, a very early adult charge is a powerful indication of subsequent crime (Berk 2018b). Figure 5 shows that the relationship with future arrests is substantially non-linear. Computed risks

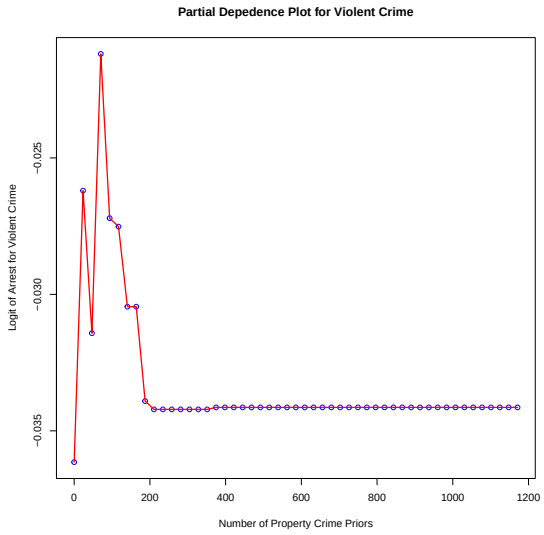

Figure 6: Partial Dependence Plot of the Association of The Number of Priors for Property Crimes with An Arrest for a Violent Crime for White Offenders Using A White-Trained Trained Algorithm

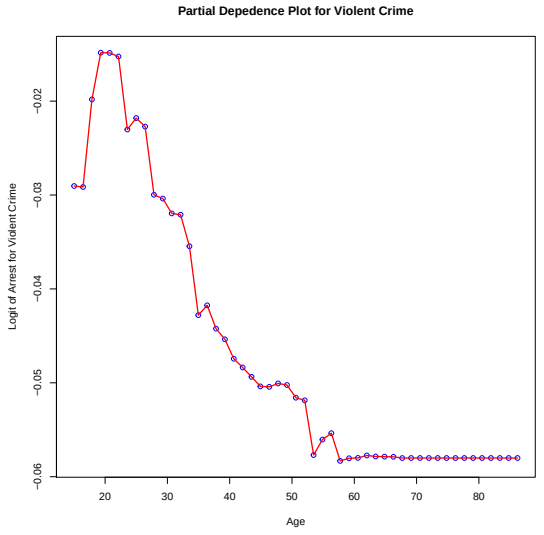

Figure 7: Partial Dependence Plot of the Association of Age with An Arrest for a Violent Crime for White Offenders Using A White-Trained Trained Algorithm

decline sharply until the age of the first adult change reaches the early 20s, levels off, and then increases somewhat in middle age. An increase in middle age is thought to be related to various kinds of family violence. The data beyond age 60 are very sparse and probably not worth interpreting.[28]

2. **The number of arrests for property crimes (in which no force is used)** – Its relationship with future arrests is thought to be positive. Figure 5, constructed using the same software, shows the partial dependence plot in which the relationship is strongly positive where the mass of the training data are located, and then turns negative for the very few case with more than 100 property priors. There are almost no cases in the training data with more than 200 property crime priors, and we wonder if those numbers were recorded in error.[29] The difference between no property priors and 25 property priors makes an important difference. Additional property priors have almost no impact on the fitted values. This underscores the point made earlier about the robustness of the fitting process and is empirically inconsistent with claims that excessive numbers of arrests are important drivers of computed risk.

3. **The age of the offender** – Age also is well-known to be a powerful predictor,

[28]Figure 5 is the partial dependence plot. We used the library *pdp* in R. Logits (i.e., logged odds) are on the vertical axis. Its formal properties are addressed in Hastie et al. (2009: section 13.10.2). The lines connecting the open circles are an interpolation, not a smoother. To reduce computation burdens, the predictor values are binned.

[29]The issue is sparse *training* data for cases with very large numbers of property priors. There is almost nothing to train on. The manner in which partial dependence plots are constructed uses the full dataset for each fitted value that is plotted; that's not the problem and illustrates why the non-parametric fitted values are robust to outliers. To learn more details of how the plotting is done when there is a large number of predictor values, the documentation for *partial* in the *pdp* library should be consulted. For Table 5, the binned, plotted points are 25 priors apart.

which typically has its peak impact in the late teens and early 20s, drops off sharply until about age 40, and then levels off (Berk, 2018b). Figure 7 is the partial dependence plot for age. Data for individuals over 60 years of age is very sparse.

Table 6: Stochastic Gradient Boosting Confusion Table form Test Data for Black Offenders Using the White-Trained Algorithm: 54% Predicted No Arrest, 30% Non-violent Arrest, 16% Predicted Violent Arrest

Observed Outcomes	No Arrest Predicted	Non-Violent Arrest Predicted	Violent Arrest Predicted	Classification Error
No Arrest	43167	16231	8353	.36
Non-Violent Arrest	16488	15917	6739	.60
Violent Arrest	5246	4563	3762	.72
Prediction Error	.33	.57	.80	

Table 6 is the test data confusion table for Black offenders constructed from the algorithm trained on White offenders; White privilege is available to White and Black offenders alike. Black offenders are a bit more likely than White offenders to be predicted to be re-arrested for a violent crime 16% to 9%. Black offenders are a bit less likely than white offenders to be predicted to experience no re-arrest whatsoever 54% to 59%. The percentages for a re-arrest for a non-violent crime are far more alike. Whether these differences represent important outcome inequalities is in some sense in the eyes of the beholder. But the over-representation of Black offenders in the violent crime class may be taken by some as evidence of algorithmic bias.[30]

[30]Proper statistical inference for stochastic gradient boosting has not yet been solved because the algorithm is adaptive, and one must take into account not just all of the regression trees there were actually constructed by the boosting algorithm, but all of the regression trees that *could* have been constructed (Berk et al., 2020). However, if one is prepared to treat the training data and the trained

For two of the three outcomes, there is a little more forecasting error for Black offenders than White offenders 33% to 30%, 57% to 51%, and 80% to 82%. Such comparisons reveal virtually no differences because of they are a product of tuning approximations. In this case, the direction the three comparisons was very stable over tunings, but the differences shown are on the high side. Still, some stakeholder could object; there might be a lack of predictive parity.

With the potential concerns about equality of outcome and predictive parity, computing risk using an algorithm trained on Whites is by itself insufficient. But a very important statistical point has been underscored: *whatever the bias, it cannot be blamed on the algorithm.* There can be no argument about this because both White offenders and Black offenders are being treated by the algorithm as if they are White. If significant unfairness remains, it *must* come from the data. Black offenders and White offenders *must* be bringing different data distributions to the risk algorithm. This realization can help change the focus of risk assessment criticisms and perhaps shift the discussion away from technical to substantive concerns. Such issues as "over-policing" and inflated arrest records come to the fore. We are back at the top of Figure 1.

---

algorithmic structure as fixed, one can with test data compute legitimate confidence intervals and statistical tests using a non-parametric bootstrap. Because the number of test observations here is very large, one easily rejects the null hypothesis of no difference at less than the .001 level for all of the White/Black comparisons, even all trivial ones. Given all of the caveats, statistical tests probably are not instructive.

Discounting Prior Arrests

If predictors such as the number of prior property arrests are important drivers of risk, and if they unfairly accumulate more rapidly for African-Americans, an obvious option is to discount the arrest counts. Recall, that arrests for relatively minor offenses and juvenile arrests already had been discarded. We sought, therefore, to construct results that might be more politically acceptable by discounting Black offenders' priors for serious crimes when the risks for re-arrests were computed. As before, we used the algorithm trained on white offenders only and applied it separately to test data from White and Black offenders. Only the test data for Black offenders was altered.

Some might claim that the number of prior arrests for Black offenders, even for felonies and violent crimes, is too large by a factor of 2 or more. With no theoretical or policy guidance, we employed for Black offenders the square root of the number of test data priors for the variety of serious priors included as predictors. Our intent was to "pull in" the right tails.

Table 7 shows the confusion table that results. The square root adjustment to the numbers of priors does not improve fairness. Moreover, prediction error had increased for all three outcome classes. Probably the most important change is that the proportion of those projected to be re-arrested for a violent crime has increased from 16% to 21%, and the gap between Black offenders and White offenders for equality of outcome has increased substantially.

The non-linear transformation is probably the culprit. The algorithm was trained on the number of priors. The test data were altered as a non-linear transformation

of those counts. Unless the algorithm arrived at close approximations of square root transformations, the likely result is a reduction in performance.

Perhaps bias makes a difference primarily for small prior counts. If it is the first few priors that really matter for risk forecasts, the prior distributions could be discounted at the low end. We divided by 2 and then recoded prior counts of 1 to 0. At arraignment, this turned many Blacks into first offenders, which changed the confusion table in a manner much like Table 7. Fairness was not improved. Although these transformations are hardly definitive, risk assessment unfairness may not derive from biases in prior record, despite the concerns of some stakeholders. Moreover, thinking about the importance of *base rates* in the earlier mathematical results, provides a significant, alternative clue.

Table 7: Adjusted Gradient Boosting Confusion Table from Test Data for Black Offenders Using the White-Trained Algorithm: 53% Predicted No Arrest, 26% Non-violent Arrest, 21% Predicted Violent Arrest

Observed Outcomes	No Arrest Predicted	Non-Violent Arrest Predicted	Violent Arrest Predicted	Classification Error
No Arrest	41927	15182	11841	.30
Non-Violent Arrest	17588	12100	9414	.70
Violent Arrest	5485	3800	4129	.69
Prediction Error	.36	.61	.84	

Altering the Base Rates

Another reason why Black offenders might have greater computed risks for violent crime re-arrests is base rate differences between Blacks and Whites. Recall that in Table 5, 9% of White offenders were in fact re-arrested for a violent crime after an

arraignment release, and in Table 6, 16% of Black offenders were in fact re-arrested for a violent crime after an arraignment release. Although small as the absolute value of the difference, their ratio seemed more notable. Disparities in base rates are known to cascade through a confusion table potentially producing several different kinds of unfairness (Kleinberg et al., 2017; Chouldechova, 2017).[31]

It is easy to change training data base rates for re-arrests for violent crime. One merely weights the training data so that post-release arrests of Black offenders for violent crimes are effectively reduced relative to post-release arrests of White offenders for violent crimes. We have undertaken such exercises elsewhere (Berk, 2019; Elzarka, 2019) showing that changing bases rates can dramatically alter the results and impact fairness.

In this case, we proceeded somewhat indirectly to reduce concerns about inequality of treatment. We ran the stochastic gradient boosting algorithm again on whites alone using weighting to discount the importance of post-release arrests for violent crime. Consequently, the algorithm did not work nearly as hard to fit post-release arrests for a violent crimes, regardless of the background of the White offender. That is, all offenders in the White training data who were more likely to be re-arrested for a violent crime, got a break. Then, because Black offenders empirically were more likely than White offenders to be re-arrested for violent crimes when risks were

[31]Black base rates for violent crimes may well be under-estimated because police clearance rates in disadvantaged neighborhoods are well known generally to be lower than in other neighborhoods, even for crimes like homicide (Lowery, 2019). Part of the explanation is that the kinds of crimes and the attributes of perpetrators create more challenges in some neighborhoods than others. For example, in some neighborhoods potential witnesses are more likely to fear for their safety and be less inclined to come forward. The mix of crimes matters too. Homicides associated with intimate partner violence, for instance, automatically define "a person of interest," who usually is easily found. Drive-by shootings typically are more challenging to solve.

computed separately for Blacks and Whites using test data, the *forecasts* of violent crime re-arrests would be more alike. In other words, all offenders were subjected to the same risk algorithm trained on White offenders and with the data weighted to discount the importance of post-release arrests for violent crimes. The indirect effect would be to make the base rates for violent crime more alike for Black offenders and White offenders because Black offenders had higher base rates to begin with.

Table 8 is the resulting confusion table for White offenders constructed from test data. Table 9 is the resulting confusion table for Black offenders constructed from test data. No adjustments were made, as we did before, for the numbers of priors for more serious crimes; interest centered on the impact of adjusted base rates.

Table 8: Stochastic Gradient Boosting Confusion Table from Test Data for White Offenders With Violent Crime Re-Arrests Down-weighted Using White-Trained Algorithm: 58% Predicted No Arrest, 42% Non-violent Arrest, 3% Predicted Violent Arrest

Observed Outcomes	No Arrest Predicted	Non-Violent Arrest Predicted	Violent Arrest Predicted	Classification Error
No Arrest	23882	11027	719	.33
Non-Violent Arrest	8131	11915	449	.42
Violent Arrest	1687	2209	375	.92
Prediction Error	.30	.53	.77	

The differences between the computed risk distributions are much smaller. For White offenders compared to Black offenders and from no re-arrest to a re-arrest for a violent crime respectively, one has 58% to 51%, 42% to 44%, and most important, 3% to 5%. Down-weighting post-release arrests for crimes of violence for all offenders altered the equality of outcome as hoped, although there may still be concerns about

Table 9: Stochastic Gradient Boosting Confusion Table from Test Data for Black Offenders With Violent Crime Re-Arrests Down-weighted Using White-Trained Algorithm: 51% Predicted No Arrest, 44% Non-violent Arrest, 5% Predicted Violent Arrest

Observed Outcomes	No Arrest Predicted	Non-Violent Arrest Predicted	Violent Arrest Predicted	Classification Error
No Arrest	41371	23664	2836	.39
Non-Violent Arrest	15182	21865	2061	.44
Violent Arrest	5315	6893	1279	.91
Prediction Error	.34	.58	.79	

racial disparities for those predicted not to be re-arrested. And predictive parity does not seems to be adversely affected. For Whites compared to Blacks, predictive error is 30% to 34% for no arrest, 53% to 58% for arrest for a non-violent crime, and 77% to 79% for a violent crime.

By these criteria, the results may well be more politically acceptable than the original results in which both White and Black offenders had their risk scores computed from test data using the white-trained algorithm. Now, all offenders are sentenced is if they were white, conditional on White and Black offenders having more similar bases rates for violent crime re-arrests. With further tuning, different cost ratios, and other adjustments for different base rates, it would be possible to do even better. For example, the figures for no arrest could be made more alike too. We doubt, however, that perfect equality is possible in the real world of criminal justice practice. But once again, the baseline is current practice. The goal is improvement.[32]

[32]If one is prepared to leave behind the real world of criminal justice practice, a perfect form equality is easily obtained. One can simply flip a coin to determine who is released at arraignment. Whatever their background, all offenders have the exact same chance of being released. Inequality of outcome and unequal treatment have been eliminated. Although classification error and prediction error can still differ for White offenders compared to black offenders, neither kind of error affects

Summary and Conclusions

We began with conventional practice. The boosting algorithm was trained on Black and White offenders together, and possible racial differences were examined using test data separately for Black offenders and White offenders. There was no compelling evidence for racial unfairness and accuracy was improved compared to predictions from the marginal distribution of the response. Unfairness apparently is not inevitable, at least when base rates across groups are alike. As a political matter, however, the empirical results might not carry the day. Nothing was explicitly done to counter strongly held beliefs that racial bias is endemic.

We then illustrated with real data analyses methods by which one can arrive at risk assessment results that perhaps are politically acceptable. We began by training a stochastic gradient boosting algorithm only on White offenders. Risk scores were then computed with test data separately from Black offenders and White offenders using the white-trained algorithm. One might argue that as a conceptual matter, treating Black offenders as if there were White provides protection against racially biased algorithms.

Some might find those results sufficiently fair. Others would be troubled by the greater fraction of Black offenders than White offenders, who were forecasted to be re-arrested for a violent crime. One possible explanation was that because of "biased" policing or "over-policing," Blacks offenders inappropriately had longer prior records.

actions any longer (i.e., the last step in Figure 1), and in that sense, are irrelevant. The price is numerous detention and release mistakes because by chance, many low risk offenders will be detained and many high risk offenders will be released. One has given up on accuracy. And if one gives up on accuracy, two closely related fair approaches are to release no one or to release everyone.

Those longer records increased the computed risks of re-arrests for crimes of violence.

We addressed those concerns in three ways. First, we had earlier discarded priors for "petty offenses" in which police could exercise considerable discretion. They played no role in how the algorithm was trained for any of the analyses and cannot be blamed for the results. In the same fashion, we also discarded prior arrests as juveniles because carrying forward the "bad decisions" by adolescents has been controversial (Loeffler and Grunwald, 2015).

Second, beginning with the white-trained algorithm, we transformed all prior counts for the Black test data by computing their square roots. The non-linear transformations changed the confusion table dramatically, and the results were less fair. The story basically was the same when we divided the prior counts by 2 and recoded all 1's to 0s.[33]

Third, differences in base rates are well known be a potential source of unfairness in criminal justice applications. Using weighting, we discounted in the White training data all re-arrests for violent crime, anticipating that implicitly the base rates by race would become more alike. Black offenders would benefit more than White offenders when re-arrests for violent crime were made relatively less common for all offenders. The over-representation of Black offenders projected to be re-arrested for crimes of violence was substantially reduced, and the proportion of Black offenders projected to be re-arrested for a violent crime was more like the proportion of White offenders projected to be re-arrested for a violent crime.[34]

[33]Generally, boosting will find the transformations of predictors that optimize the fit. Figures 5, 6, and 7 are illustrations. It is very unlikely that any of the variables for priors were algorithmically transformed by a close approximation of their square root.

[34]It might seem that similar options could be applied to the test data. But test data base rates do

There are surely no guarantees that politically acceptable results will be obtained with other data, and perhaps our remarkably fair initial results were an aberration. Regardless, we illustrated some potential, remedial strategies for risk assessments initially criticized as unfair.

1. Exclude predictors that arguably are especially vulnerable to racial bias. Good candidates are prior arrests for crimes in which police can exercise extensive discretion. Perhaps exclude juvenile offenses as well.

2. Train the risk algorithm with training data from the most privileged group. Then, compute risk for members of all groups separately from test data using the results from the algorithm trained on the most privileged group. These two steps alone my yield sufficient fairness. In addition, there can be no longer legitimate claims that the algorithm itself is unfair.

3. Consider discounting the impact of priors in test data for the less privileged groups. But a sensible transformation must be determined that do not create new problems.

4. Also, consider retraining the algorithm on data from the privileged group discounting re-arrest for crimes thought to foster unfair forecasts. The discounting is done for all groups at once when risk is computed, but the indirect impact should be to affect the most groups with the higher base rates.

not figure in how the fitted values are obtained. One could randomly discard some of the re-arrests for Black offenders in the test data, but when time came for real forecasts, there would be no known re-arrests to discard.

In practice, various combinations of these potential strategies would need to be employed, and the resulting confusion tables carefully examined by stakeholders. Frequently, there will be analytical tools that can help document the tradeoffs (Berk et al., 2017; Kearns et al., 2018b). Many approaches are likely to be widely unacceptable, which can simplify the discussion. Of the remaining results, several will have the most support and often be so similar that there is not much to argue about; all may be politically acceptable, and any could break the stalemate. Finding a single confusion table that dramatically dominates all others is probably unrealistic.

We emphasize that within our approach, a politically acceptable risk tool is one for which they can be sufficient support. There is no deeper meaning and no requirement that the risk tool be objectively fair. Both issues introduce difficult philosophical matters not easily addressed, let alone resolved. For example, a politically acceptable reform may in theory optimize the overall utility of a political jurisdiction, but that says nothing about inequality. "Bentham's concern – and that of utilitarianism in general – was with the total utility of a community. The focus was on the total sum of utilities, irrespective of the distribution of that total, and in this we can see some blindness of considerable ethical and political concern" (Sen, 2018: 10).

Links to the current jurisprudence remain to be made. We lack the necessary expertise and here too, the issues are challenging (Hyatt et al., 2011; Tonrey, 2014; Starr, 2014; Goel et al., 2016; Hamilton, 2016; Huq, 2019; Mayson, 2019). One important consideration is that by training a risk algorithm only on the more privileged groups, no group is made worse off, and one or more groups are made better off. We have a form of Pareto improvement. However, until the legal implications

are clarified, implementation may be an invitation for litigation.

Finally, the impact of each our strategies must be evaluated for their effect on *victims*. In particular, any approach that discounts prior record and/or re-arrests for certain kinds of offenses, discounts the harm to victims of those crimes. Often those victims will be disproportionately from disadvantaged neighborhoods. In 2018, there were 351 homicides in the city from which our data were collected. 92% of the victims were African-American (McSwan, 2019). Is it fair to discount their deaths?

References

Arrow, K.J. (1951) *Social Choice and Individual Values*, first edition, New York: Wiley.

Berk, R.A. (2018a) "A Primer on Fairness in Criminal Justice Risk Assessments." *Translational Criminology* Issue 15: 8–11.

Berk, R.A. (2018b) *Machine Learning Forecasts of Risk in Criminal Justice Settings.* New York Springer.

Berk, R.A. (2019) "Accuracy and Fairness for Juvenile Justice Risks Assessments." *Journal of Empirical Legal Studies*, published online February, 2019 (DOI 10.1111/jels.12206)

Berk, R.A., Sorenson, S.B., and Barnes, G. (2016) "Forecasting Domestic Violence A Machine Learning Approach to Help Inform Arraignment Decisions." *Journal of Empirical Legal Studies* 13(1) 94–115.

Berk, R.A., Barnes, G., Alhman, L., and Kurtz, E. (2010) "When Second Best Is Good Enough A Comparison Between a True Experiment and a Regression Discontinuity Quasi-Experiment." *Journal of Experimental Criminology* 6(2) 191–208.

Berk, R.A., and J. Bleich (2014) "Forecast Violence to Inform Sentencing Decisions." *Journal of Quantitative Criminology* 30 79–96.

Berk, R.A., and Hyatt, J. (2015) "Machine Learning Forecasts of Risk to Inform Sentencing Decisions." *The Federal Sentencing Reporter* 27(4) 222 – 228.

Berk, R.A., Heidari, H., Jabbari, Kearns, M., Morganstern, J., Neel, S., and Roth, A. (2017) "A Convex Framework for Fair Regression." *arXiv" 1706.02409v1 [cs.LG]*

Berk, R.A., Heirdari, H., Jabbari, S., Kearns, M., and Roth, A. (2018) "Fairness in Criminal Justice Risk Assessments The State of the Art." *Sociological Methods and Research*, first published July 2nd, 2018, http //journals.sagepub.com/doi /10.1177/0049124118782533.

Berk, R.A., Buja, A., Olson, M, and Ouss, A. (2020) "Using Recursive Partitioning to Find and Estimate Heterogenous Treatment Effects In Randomized Clinical Trials." *Journal of Experimental Criminology*, forthcoming, 2020.

Blinder, A.S. (1973). "Wage Discrimination Reduced Form and Structural Estimations." *Journal of Human Resources* 8 436–455.

Bloom, S. L. (2014). "The Impact of Trauma on Development and Well-Being." In K. G. Ginsburg & S. B. Kinsman (Eds.), *Reaching Teens - Wisdom From Adolescent Medicine.* Elk Grove Village, IL American Academy of Pediatrics.

Boonin, D. (2011) *Should Race Matter?* Cambridge Cambridge University Press.

Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). "A Training Algorithm for Optimal Margin Classifiers." Proceedings of the fifth annual workshop on Computational learning theory ? COLT '92: 144–150.

Breiman, L. (2001a) "Statistical Modeling Two Cultures" (with discussion). *Statistical Science* 16 199–231.

Breiman, L. (2001b) "Random Forests." *Machine Learning* 45: 5–32.

Chouldechova, A. (2017) "Fair Prediction With Disparate Impact A Study of Bias in Recidivism Prediction Instruments." arXiv 1610.075254v1

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Hug, A. (2017) "Algorithmic Decision Making and Cost of Fairness." *KDD '17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Corbett-Davies, S. and Goel, S. (2018) The Measure and Mismeasure of Fairness A Critical Review of Fair Machine Learning. 35th International Conference on Machine Learning (ICML 2018).

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science* 243(4899) 1668–1674.

Doleac, J, and Stevenson, M. (2106) "Are Criminal Justice Risk Assessment Scores Racist?" Brookings Institute. https //www.brookings.edu/blog/up-front/2016/08/22/are-criminal-risk-assessment-scores-racist/

Donovan, T., Smith, D.A., Osborn, T and Mooney, C.Z. (2014) *State & Local Politics: Institutions and Reform*, fourth edition. Boston: Cengage Learning.

Elzarka, A. (2019) "Establishing Fairness in Algorithms." Thesis in Data Science, presented to the faculties of the University of Pennsylvania in partial fulfillment of the requirements for the degree of master of science in engineering.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., and Venkatasubrtamanian, S. (2015) "Certifying and Removing Disparate Impact." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 259 – 268. //

Friedler, S.A., Scheidegger, C., and Venkatasubramanian, S. (2016) "On The (Im)possibility of Fairness)." axXiv1609.07236v1 [cs.CY].

Ferguson, A.G. (2017) *The Rise of Big Data Policing Surveillance, Race, and the Future of Law Enforcement* New York New York University Press.

Friedman, J. H. (2001) "Greedy Function Approximation A Gradient Boosting Machine. " *The Annals of Statistics* 29 1189–1232.

Friedman, J. H. (2002) Stochastic Gradient Boosting. *Computational Statistics and Data Analysis* 38 367–378.

Goel, S., Rao, J.H., and Shroff, R. (2016) Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics* 10(1) 365 – 394.

Goel, S., Shroff, R., Skeem, J. L. and Slobogin, C. (2018) "The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment." Available at SSRN: https://ssrn.com/abstract=3306723 or http://dx.doi.org/10.2139/ssrn.3306723

Grogger, J. and Ridgeway, G. (2006). "Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness." *Journal of the American Statistical Association* 101(475) 878–887.

Guay, J. P., and Parent, G. (2018). "Broken Legs, Clinical Overrides, and Recidivism Risk: An Analysis of Decisions to Adjust Risk Levels With the LS/CMI." *Criminal Justice and Behavior* 45(1): 82–100.

Harcourt, B.W. (2007) *Against Prediction Profiling, Policing, and Punishing in an Actuarial Age*. Chicago, University of Chicago Press.

Hardt, M., Price, E., Srebro, N. (2016) "Equality of Opportunity in Supervised Learning." In D.D. Lee, Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett (eds.) *Equality of Opportunity in Supervised Learning*. Advances in Neural Information Processing Systems 29 Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, (pp.3315 – 3323).

Hamilton, M. (2016) "Risk-Needs Assessment Constitutional and Ethical Challenges." *American Criminal Law Review* 52(2) 231 – 292.

Hastie, T., Tibshirani, R., and J. Friedman (2009) *The Elements of Statistical Learning*. New York Springer.

Heissel, J.A., Sharkey, P.T., Torrats-Espinosa, G., Grant, K., and Adam, E.K. (2017) Violence and vigilance the acute effects of community violent crime

on sleep and cortisol. *Child Development*, published online, doi 10.1111/cdev.12889.

Huq, A.Z. (2019) "Racial Equality in Algorithmic Criminal Justice." *Duke Law Journal* 68 (6), 1043–1134.

Hyatt, J.M., Chanenson, L. and Bergstrom, M.H. (2011) Reform in motion the promise and profiles of incorporating risk assessments and cost-benefit analysis into Pennsylvania Sentencing. *Duquesne Law Review* 49(4) 707–749.

Jung, J., Concannon, C., Shroff, R., Goel, S., and Goldstein, D. G. (2017). "Simple Rules for Complex Decisions." Unpublished manuscript available at: https://arxiv.org/pdf/1702.

Kamiran, F., and Calders, T. (2009) "Classifying Without Discrimination." *2009 2nd International Conference on Computer, Control and Communication*, IC4 2009.

Kamiran, F., and Calders, T. (2012) "Data Preprocessing Techniques for Classification Without Discrimination." *Knowledge Information Systems* 33 1 - 33.

Kamiran, F., Karim, A., and Zhang, X. (2012) "Decision Theory for Discrimination-Aware Classification." IEEE 12th International Conference on Data Mining.

Kamishima, T., Akaho, S., and Sakuma, J. (2011) "Fairness-aware Learning Through a Regularization Approach." Proceedings of the 3rd IEEE International Workshop on Privacy Aspects of Data Mining.

Kearns, M., Neel, S., Rith, A, and Wu, Z. (2018a) Preventing Fairness Gerrymandering Auditing and Learning Subgroup Fairness." arXiv 1711.05144v4 [cs.LG].

Kearns, M., Neel, S., Roth, A, and Wu, Z. (2018b) An empirical study of rich subgroup fairness for machine learning. asXiv 1808.08166v1 [cs.LG]

Kitagawa, E. (1995)"Components of a Difference Between Two Rates." *Journal of the American Statistical Association* 50 1168–1194.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017a) "Inherent Tradeoffs in the Fair Determination of Risk Scores." Proc. 8th Conference on Innovations in Theoretical Computer Science (ITCS).

Kleinberg, J., Lakkaraju, H., Loskovec, J., Ludwig, J, and Mullainathan, S. (2017b) Human decisions and machine predictions.*Quarterly Journal of Economics* 133(1) 237 293.

Loeffler, C., and Grunwald, B. (2105) "Decriminalizing Delinquency: An Estimate of the Effects of Raising the Age of Majority for a Juvenile Court." *Journal of Legal Studies* 44(2): 361–388.

Lowery, W., Kelly, K., Melnick, T & S.Rich (2018) "Where Killings Go Unsolved." Washington Post, June 6, 2018. https //www.washingtonpost.com/graphics/2018/in murders-go-unsolved/

Mayson, S.G., (2019) "Bias In, Bias Out." *The Yale Law Journal* 128 2218–2300.

Meehl, P.E. (1954) *Clinical versus Statistical Predictions* Minneapolis: University of Minnesota Press.

McSwain, W.M. (2019) "Police Deserve DA's Support." Philadelphia Inquirer, May 12 12, 2019, page C4.

Mease, D., Wyner, A.J., and Buja, A. (2007) Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research* 8 409–439.

Mullainathan, S. (2019) "Biased Algorithms Are Easier to Fix Than Biased People." NewYork Times: https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html.

Oaxaca, R. (1973). "Male-Female Wage Deferential in Urban Labor Markets." *International Economics Review* 14 (3) 693–709.

Rawls, J. (2001) *Justice as Fairness A Restatement*, second edition. Cambridge Harvard University Press.

Ridgeway, G. (2006) "Assessing the effect of race bias in post-traffic stop outcomes using propensity scores." *Journal of Quantitative Criminology* 22(1) 1–29.

Ridgeway, G. (2007) "Generalized Boosted Models A Guide to the gbm Package." cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.

Ridgeway, G (2016) "Officer Risk Factors Associated with Police Shootings A Matched Case-Control Study." *Statistics and Public Policy* 3(1) 1–6.

Ridgeway, G. and MacDonald, J.M. (2014). "A Method for Internal Benchmarking of Criminal Justice System Performance." *Crime & Delinquency* 60(1) 145–162.

Royal Statistical Society (2018) *Algorithms in the Justice System: Some Statistical Issues.* London, EC1, United Kingdom.

Sen, A. (2009) *The Idea of Justice.* Cambridge Harvard University Press.

Sen, A. (2018) *Collective Choice and Social Welfare* Cambridge: Harvard university Press

Speicher, T., Heidari, H., Grgic-Hlaca, N., Krishna, G., Gummadi, G., Singla, A., Weller, A., Zafar, M.B. (2018) "A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices." KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2018, London, United Kingdom.

item Sowell, T. (2002) *The Quest for Cosmic Justice.* New York Free Press.

Starr, S.B. (2014) "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review* 66 803 – 872.

Tonry, M. (2014) "Legal and Ethical Issues in The Prediction of Recidivism." *Federal Sentencing Reporter* 26(3) 167 – 176.