Algorithms as Prosecutors: Lowering Rearrest Rates Without Disparate Impacts and Identifying Defendant Characteristics ‘Noisy’ to Human Decision-Makers

Daniel Amaranto, Elliott Ash, Daniel L. Chen, Lisa Ren, Caroline Roper

January 28, 2018

Abstract

This paper investigates how machine learning might bring clarity to human decisions made during the criminal justice process. Our data comes from all cases at the New Orleans District Attorney’s office for the years 1988-1999. We exploit random assignment of prosecutors, prosecutorial discretion, and heterogeneity across prosecutors in charge rates to compare prediction models to judicial decision makers. Our model of defendant rearrest, trained using defendant and offense characteristics, selects higher-risk individuals to prosecute than its human counterparts did. In particular: given a set charge rate, our model would reduce rearrest rates by five to nine percentage points. This model could have several important policy implications: it might identify defendant characteristics that are particularly ‘noisy’ to prosecutors; it could suggest ways of alleviating criminal caseloads without increasing crime rates; and it might provide important insights into how a prosecutor’s background relates to the quality and nature of their charging decisions.

1 Introduction

A widely held view is that the U.S. criminal justice system incarcerates too many and often does so unfairly (e.g., Fagan and Ash, 2017). Experts have identified certain major factors underlying the system's purported inefficiencies and inequities; these include, in particular, the aggressiveness of prosecutors and plea-bargaining practices (Gopnik, 2017). Previous work has shown that prosecutors play a major role in criminal justice outcomes. Wilson et al (2011) estimate that for every case seen by a judge, there are 15 cases seen by a prosecutor. Rehavi and Starr (2014) shows that prosecutor charge type can explain most of the racial disparity in sentencing. According to Pfaff (2016), half of the increase in felony filings between 1990 and 2010 were due to misdemeanors being charged by prosecutors as felonies.

After a person is arrested and before a trial begins, prosecutors (screeners) can decide to either accept those charges and proceed to a trial or to drop them. This is a quasi-judicial role, but it has not gotten much attention, despite its role in fairness and efficiency of the system, due to lack of data (Miller and Wright, 2002). Chen (2017) shows that prosecutorial screening substantially contributes to the racial gap in criminal justice outcomes, and that there are significant interactions between racial preferences of prosecutors and judges. Given the critical role of prosecutorial decisions on criminal-justice outcomes, we aim to augment the understanding of screening decisions made by prosecutors.

In order to assess whether or not the decision to drop charges was made correctly, we use rearrest as our target; that is, if an individual who had charges dropped enters the arrest registry again within a certain time frame, we consider the screen decision to have been wrong. To optimize this prediction problem we use gradient boosted trees, a forward stagewise additive modeling algorithm that averages decision trees that are sequentially improved. After optimizing the model, we employed techniques described by Kleinberg et al. (2017) to assess its performance compared to screeners. A reduction in rearrest rate model by the model would allow us to critique the way that screeners select defendants to charge.

2 Background and Data

The information in this section on the setting, data, and construction of variables come from Miller and Wright (2002), unless otherwise noted.

Local prosecutors, in offices generally headed by an elected District Attorney, are the attorneys who represent the government in taking individuals and corporations to court for criminal offenses. Harry Connick, the District Attorney in New Orleans from 1973–2003 developed a process that heavily emphasized the screening of cases prior to accepting them for prosecution. The goal was to minimize the number of plea-bargains and improve the quality of cases taken to court.

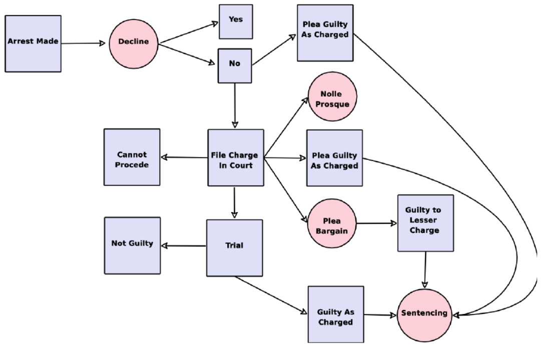
Connick perceived that the large quantity of cases handled by a sizeable staff with high turnover rates would create continual pressure to resort to plea bargaining and frustrate this arrangement. To create countervailing forces, he created a database that enabled him to continually monitor and supervise the casework being done.

After finishing their investigation, police officers build a case folder that is filed with the magistrate. The first contact for the case file with the District Attorney's staff is the Magistrate Section where the least experienced (less than half-a-year in the District Attorney's office) attorneys work. The ADA from the Magistrate Section represents the state at the first appearance and bail hearing before the magistrate. The public defender or retained defense attorneys are also present for this hearing. Following this activity in the Magistrate Division, the case file moves to the Screening Section of the District Attorney's office.

Inside the Screening Segment, major cases like homicide or rape are given to screeners with additional proficiency. Drug cases and a couple of other high frequency charges go to the Expedited Screening section. Normal cases go to the screening attorney drawing cases for that day.[1] The screener surveys the case file, speaks with

[1] First, information on the case is received by a set of intake attorneys who routinely process the cases by collecting the potential defendant's rap sheet and other information. Then, once it has been processed, a clerk then receives the file and assigns it to the screening attorney. The 'duty DA' for the day (rotating duty) handles everything that arrives on a given day except for the major crimes (such as homicide) assigned to a specialist. For special crimes, the clerk just assigns the case to whoever is next up in the rotation (for example, a particular homicide prosecutor receives two out of every five homicide cases, because there are two and a half homicide screeners).

Figure 1: New Orleans District Attorney Case System



key witnesses and the victims, and assesses the strength of the case. If there is a specific issue of concern, the screening lawyer will talk directly with the relevant police officer to review it. There is a strong norm that the Screening Lawyer will settle make a decision on prosecuting within ten days of receiving the case file. The District Attorney's office initiated an assortment of measures to guarantee sensible consistency in screening choices.

Managers survey all charge rejections. The DA's policy also promotes prosecution and acceptance of certain classes charges, most especially domestic violence charges. For the most severe charges (e.g. rape and homicide), the DA's office held "charge conferences" with both police officers and higher-ranking prosecutors to discuss the case and possible charges.

2.1 Overview

We obtained the case management system database for the New Orleans Parish District Attorney's office. The data includes all cases for the years 1988 through 1999. Appendix 7.1 provides details for how the data were cleaned and merged.

The database provides detailed information on 280,000 cases involving 145,000 unique defendants. For each defendant, we have a large set of personal characteristics, including age, gender, race, ethnicity, and home address. For each case, we have the list of charges, as well as the prosecutor and judge assigned to the case. The data also contains detailed information on victims, witnesses, defense attorneys, police officers, and many other topics.

This dataset has some major advantages over other datasets in the criminal justice literature. The National Crime Victimization Survey only has data on victims. The Uniform Crime Report only has data on reported crimes. Fryer (2016) only has data on arrests. Fagan and Ash (2017) have separate datasets on municipal court decisions and policing, but they are not linkable (same for Mueller-Smith 2014). Other data sets are not at the case level so cannot be used to examine the choices of individual officials (e.g. Ash et al., 2017). Because this dataset has names, birthdates, and social security numbers, it can be linked to a range of administrative datasets. For example, Chen and Phillippe (2017) show that defendants who are sent to trial at the prosecutor stage are *more* likely to vote in subsequent elections, suggesting a politicizing effect of criminal-justice-system involvement. Arguably the closest data set to this one is Norris (2017), who exploits a two-step judicial process (with a reviewing judge) to examine judicial disagreement and errors.

2.2 Case Assignment

A feature of our empirical context is that cases are randomly assigned to prosecutors (and to judges, if the cases are not screened out). Once at the courts, cases are assigned by a clerk on duty in an arbitrary rotation. In addition, felony cases are randomly scheduled across dates to prevent the prosecutor from choosing a specific trial judge.

Table 1 provides the results of a battery of randomization checks. The left column gives our list of pre–determined defendant-level variables. The middle column shows that few of these variables are correlated with the leniency of the assigned prosecutor. The right column, however, shows that these variables are systematically related with the tendency to be screened in to the system. The relatively weak relationships in the middle column are consistent with randomization, with any observed effects being

Table 1: Testing for Random Assignment of Cases to Prosecutors

Dependent Variable:	**Prosecutor Leniency**		**Screened In (Case Accepted)**	
Pre-determined characteristics	coef.	(s.e.)	coef.	(s.e.)
Days from Police Report to Screening Date	-0.00959	(5.691)	-2.186**	(0.678)
Days from Arrest to Screening Date	5.243	(13.80)	-16.95**	(3.549)
Predicted Screen	-0.000335	(0.00230)	0.104*	(0.0481)
Missing Phone Number	-0.000826	(0.00676)	-0.00530	(0.00327)
Days between Police Report and Arrest	12.31	(9.820)	-10.08**	(1.633)
Detained at End of Arrest Proceedings	0.00205	(0.00714)	0.0284*	(0.0139)
Detained at Start of Arrest Proceedings	0.00432	(0.00655)	0.0333+	(0.0179)
Height	0.172	(0.107)	0.134**	(0.0389)
Male	0.0217	(0.0163)	0.0267**	(0.00639)
Weight	1.071	(1.108)	0.435	(0.339)
Birth Year	1.473	(1.010)	-0.550*	(0.266)
Out-of-state Drivers License	-0.00581	(0.00441)	-0.00291*	(0.00145)
Born out-of-state	-0.0110	(0.00942)	0.00848	(0.0103)
Race coded as Black	0.000840	(0.0121)	0.00545	(0.00348)
Race coded as Asian	-0.00294+	(0.00164)	-0.00135**	(0.000462)
Race coded as Hispanic	-0.000329	(0.00127)	0.000428	(0.000303)
Race coded as Indian	0.000169	(0.000310)	-0.000190*	(0.0000920)
Race coded as Negro	0.0270	(0.0224)	0.00658	(0.00504)
Race coded as Oriental	0.0000244	(0.00179)	-0.000274	(0.000621)
Race coded as Other	-0.0000924	(0.000983)	0.000571	(0.000442)
Race coded as White	-0.0247	(0.0191)	-0.0112*	(0.00434)
Skin coded as Black	-0.00111	(0.00164)	0.000488	(0.000587)
Skin coded as Brown	-0.00562	(0.00598)	0.00240	(0.00171)
Skin coded as Dark	-0.000723	(0.00239)	0.000184	(0.000354)
Skin coded as Fair	-0.00129	(0.00489)	-0.00377**	(0.00120)
Skin coded as Light Brown	0.00424	(0.00825)	-0.000217	(0.00218)
Skin coded as Light	-0.00130	(0.000850)	0.0000139	(0.000204)
Skin coded as Medium	-0.00323	(0.00241)	-0.000844+	(0.000426)
Skin coded as Olive	-0.00297	(0.00192)	-0.000239	(0.000564)
Skin coded as Sallow	0.000310*	(0.000144)	0.0000619	(0.0000554)
Skin coded as Yellow	-0.000406	(0.000355)	-0.000134	(0.000121)
Skin coded as Dark Brown	0.000285	(0.0201)	0.0173**	(0.00543)
Skin coded as Medium Brown	0.0276+	(0.0140)	-0.00741+	(0.00425)
Skin coded as Ruddy	-0.0157	(0.0173)	-0.00788*	(0.00362)
Eyes coded as Brown	0.0104	(0.0169)	0.00728**	(0.00265)
Eyes coded as Blue	-0.0120	(0.0113)	-0.00419**	(0.00157)
Eyes coded as Brown	-0.000320	(0.00195)	-0.000934	(0.000683)
Eyes coded as Green	-0.00671	(0.00476)	-0.000324	(0.000900)
Eyes coded as Hazel	0.0105	(0.00753)	-0.00187	(0.00130)
Eyes coded as Grey	-0.00187+	(0.000963)	0.0000450	(0.000205)
Hair coded as Brown	-0.00618	(0.0156)	-0.00830*	(0.00414)
Hair coded as Black	0.0235	(0.0215)	0.0129**	(0.00462)
Hair coded as Bald	-0.000223	(0.00219)	-0.000232	(0.000474)
Hair coded as Blond	-0.00534	(0.00410)	-0.00252*	(0.00106)
Hair coded as Grey	-0.0135+	(0.00800)	-0.00116	(0.000870)
Hair coded as Red	0.00220	(0.00283)	-0.000744	(0.000936)
Hair coded as Sandy	-0.000347+	(0.000177)	0.0000517	(0.0000702)
Hair coded as White	-0.0000129	(0.000200)	0.0000219	(0.0000760)

spurious and due to noise.[2]

2.3 Prediction Target

The target is a dummy variable equaling one if the defendant is rearrested within five years of the screening event.[3] This is different from Kleinberg et al. (2017), who use failure to appear in court (FTA) as the risk variable in the main analysis. In terms of a welfare analysis, the risk of re-arrest for additional crime is likely more important that the risk of FTA. Because the data ends in 1999, we only use original arrests for the years 1989 through 1994. That way, we have the full five years of subsequent arrest records to use for our re-arrest target.

2.4 Predictors

We use the following variables in our feature set: (1) date of screening disposition (SCREENING_DISP_DATE) converted to year and month indicators; (2) date of arrest (ARREST_DATE), converted to year and month indicators; (3) time between arrest and screening (ARREST_TO_SCREEN); (4) the number of days elapsed between police report date and screening date (SCREENING_DAYS); (5) the number of days elapsed between the arrest and the filing of the police report (POLICE_RPT_DAYS); (6) whether defendant was detained at the beginning of the arrest proceedings (INITIAL_DETENTION_FLAG); (7) whether defendant was detained at the end of arrest proceedings (FINAL_DETENTION_FLAG); (8) defendant race (RACE); (9) defendant gender being male (SEX); (10) defendant criminal history (CRIMINAL_FLAG); (11) whether the defendant is a habitual offender at arrest (HABITUAL_OFFENDER_FLAG); (12) whether the defendant is juvenile (JUVENILE_FLAG); (13) the manner in which the defendant was charged at arrest (CHARGE_TYPE); (14) the severity of the charge at arrest (CHARGE_CLASS); (15) total number of defendants (TOT_NUM_DEF);

[2] Predicted screen is from a linear model regressing the screener decision on pre-determined characteristics.

[3] We ran the analysis with shorter re-arrest windows with comparable results, although the model did best predicting rearrest within 5 years.

Table 2: Summary Statistics on Screening and Re-Arrest

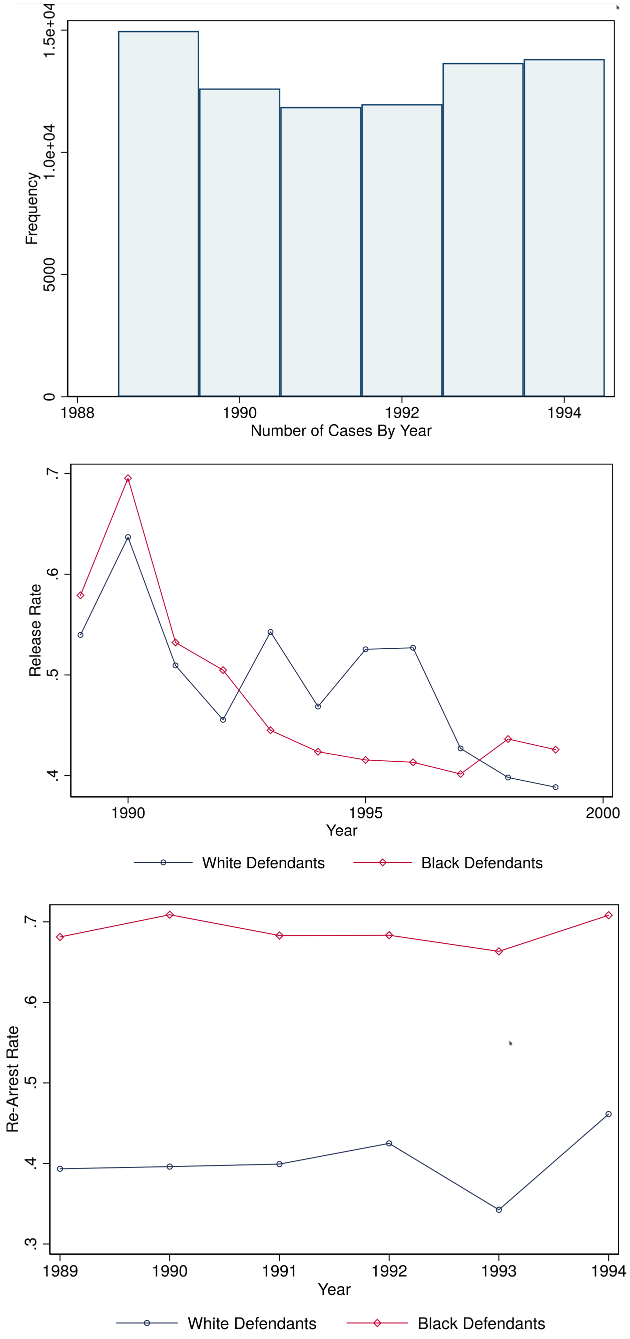


Table 3: Summary Statistics on Predictors

Variable	Mean	Std. Dev.
Binary Predictors		
CRIMINAL_FLAG	0.8	0.4
FINAL_DETENTION_FLAG	0.04	0.21
HABITUAL_OFFENDER_FLAG	0.05	0.21
INITIAL_DETENTION_FLAG	0.05	0.21
JUVENILE_FLAG	0.25	0.43
SADA_SEX	0.47	0.66
SEX	0.81	0.42
MULTIPLE_DEF_FLAG	0.11	0.31
Real-Valued Predictors		
TOT_NUM_DEF	1.18	0.72
SCREENING_DAYS	15.34	33.75
POLICE_RPT_DAYS	33.29	66.07

(16) the screening prosecutor's race (SADA_RACE); (17) the screening prosecutor's gender being male (SADA_SEX); (18) the screening prosecutor's party affiliation (SADA_PARTY); (19) days since the screening prosecutor's date of bar admission (BAR_ADMIT_DAYS); and (20) code indicating the law enforcement agency receiving credit for the arrest (ARREST_CREDIT_CODE).

Summary statistics for binary and real-valued predictors are reported in Table 3. Tabulations for categorical variables are reported in the appendix.

2.5 Data Split

We split the data into training, validation, and test sets using a split of 64/16/20. The split was stratified along the year of arrest so that the distribution of arrests over time was consistent among the training, validation, and test sets. The training and validation sample are used in Section 3. The test sample is saved for the application in Section 4.

3 Model Training and Validation

We face a straight-forward machine prediction task. Given a set of inputs, we train a model to predict the probability of rearrest. We then evaluate the model on the validation set. For a given arrestee, the model outputs a score that can be used to rank the arrestees by their predicted risk. We can then use the risk ranking to assess room for improvements in rearrest rates.

3.1 Baseline Model

The naive baseline model is a decision tree with max depth four. We included the two features we believed would be most predictive: AGE and ARREST_CLASS (severity of charge on scale from 1 to 7). The baseline model achieved 60% accuracy on the validation set and an F-score of 0.65.

3.2 Feature Selection

Next we turned to features to optimize the model. We started off with two ensemble model classes, random forests and boosted trees. We performed feature selection using the following steps:

1. Create 22 training datasets, one using all features and 21 created by removing one feature at a time.

2. Train a gradient boosted trees and a random forest model on each of the modified training data sets.

3. Evaluate each of the models to identify which feature, target, and model combination results in the highest validation F-score.

We found that the highest F-score resulted from keeping all features except Arrest Credit Code (the law enforcement agency receiving credit for the arrest).

3.3 Model Selection

After choosing a final set of features, we performed parameter optimization on two model types: random forest and gradient boosted tree models. All the models we trained use 50% as the probability threshold for each class, since the classes are balanced in our data (52% of arrestees were rearrested).

For the random forest model, we trained the model with the following hyperparameter values for a total of 54 models:

1. number of estimators (100, 300, 500)
2. maximum features ($\sqrt{\#\text{ of features}} = 9$, $\log_2(\#\text{ of features}) = 6$)
3. max depth (8, 10, 12)
4. minimum sample split (2,4,8)

We used gini index as the measure of the quality of split. The highest F-score for the random forest model was 0.7655, a significant improvement on the baseline model.

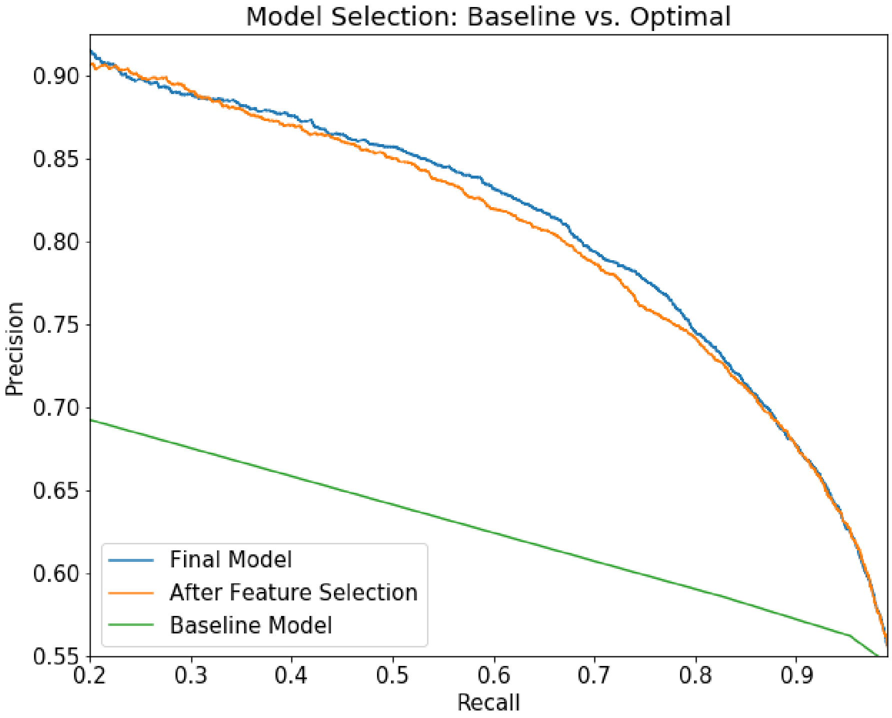
For the gradient boosted trees, we trained the model with the following values of hyperparameters for a total of 81 models:

1. number of estimators (100, 300, or 500)
2. learning rate (.05, .1, .5)
3. max depth (3, 5, 10)
4. minimum samples split (2,4,8)

All the gradient-boosted tree models use deviance as the loss function and the Friedman mean squared error as the measure of the quality of a split.

Our best-performing final gradient boosted trees model used 300 estimators, a learning rate of 0.1, a maximum depth of 5, and a minimum samples split threshold of 4. The associated F-score on the held-out validation data is 0.7703, which is better than the random forest model and .12 higher than the baseline model. Figure 2

Figure 2: Model Performance



	Predicted	Predicted	
	No Rearrest	Rearrest	
No Rearrest	3909	1387	5296
Rearrest	1260	4438	5698
	5169	5825	

provides a comparison of the final model to the baseline and illustrates the substantial improvement achieved. The confusion matrix provides additional statistics on the effectiveness of the model.

In addition, we assess the actual risk of arrest compared to the predicted risk. Using the validation set, we grouped the arrestees into quintiles by their estimated risk. We found that the predicted riskiest arrestees have higher rearrest rates. This is shown graphically in Figure 2. This suggests that the defendants who were released by a screener and predicted by our model to be risky were in fact risky. The binscatter diagrams in the bottom panel of this figure strengthen the argument. White defendants cluster around low arrest probability, while black defendants are distributed evenly.

4 Model Application: Algorithms as Prosecutors

Next we assessed the performance of our model against human screeners.[4] As shown in Figure 3, the NODA dataset provided sufficient variation in screener charge rates for us to analyze the marginal cases. The bulk of cases (roughly 60%) are seen by screeners with charge rates between 35% and 60%. We focus on this range.

4.1 Screeners are not using re-arrest risk

How the screeners rank the risk of the arrestees is unobserved. However, we can assess their implicit risk ranking from the variation in charge rates between "strict" and "lenient" screeners, by comparing the distribution of predicted risk of the arrestees charged by the "strict" and the "lenient" screeners.

Figure 4 reports this finding graphically. What we see is that the actual risk distribution amongst strict and lenient screeners differs from what we would expect to see if the screeners were releasing based on predicted risk. If screeners were

[4] This draws from the reasoning in Section 4.2 of Kleinberg et al. (2017).

Table 4: Actual Rearrest Rate, by Predicted Risk Quintile

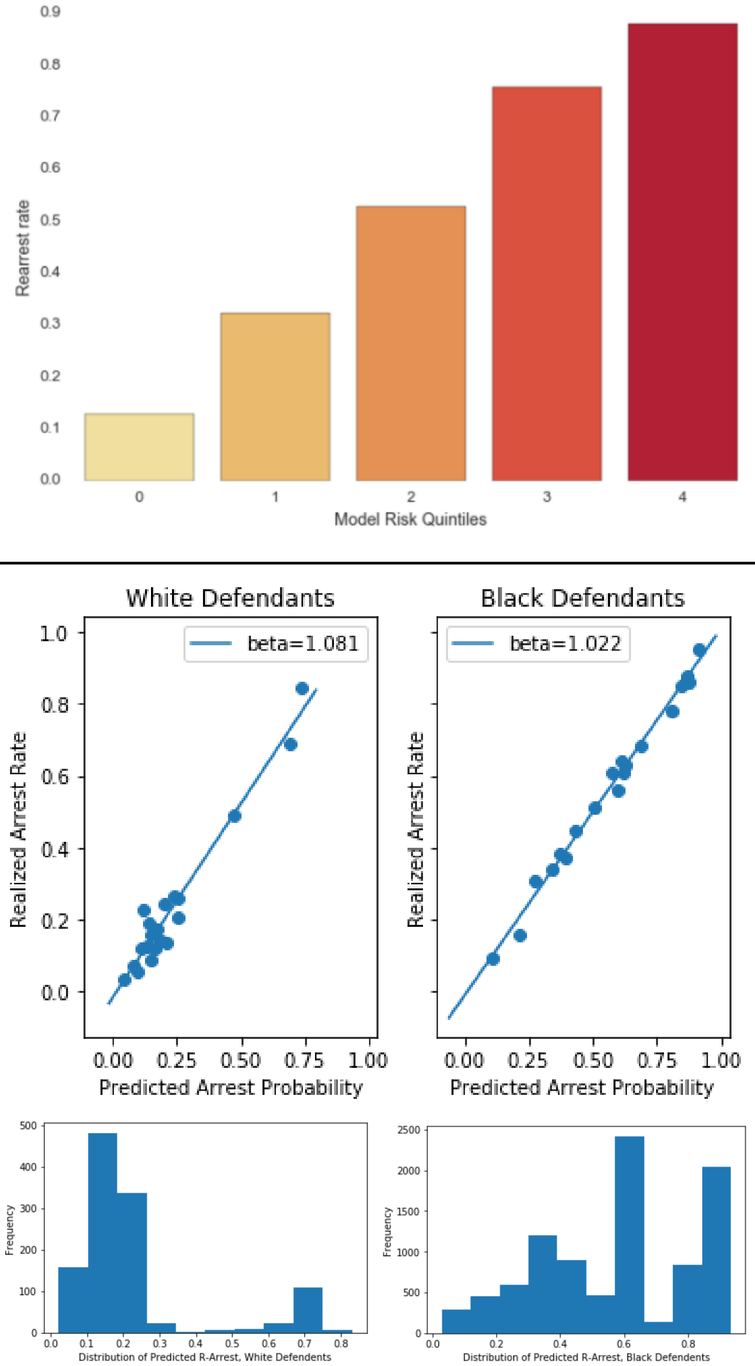
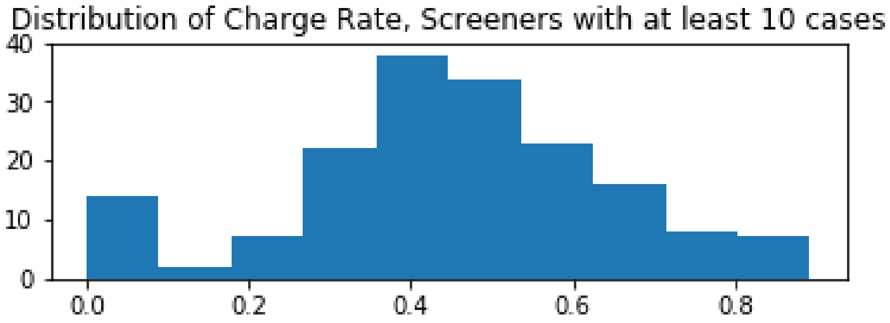


Figure 3: Distribution of Charge Rates for Screening Prosecutors



to release defendants at random, we would expect to see an even distribution of predicted risk for each set of screeners. This is similar to what we see in the actual risk composition (top panel). If they were trying to minimize re-arrest, we would see something like the middle panel, where only the lowest-risk defendants are released by the screeners with the highest charge rates. As shown in the bottom panel, there is really no relationship between charge rate and re-arrest rate. If anything, as seen in the top panel, the least-risky defendants (in green) are a shrinking share of those released.

We can also assess the performance against actual rearrest rates (rather than the risk score). In the bottom panel of Figure 5, we should observe a diagonal down-ward slope from the upper left to the lower right, if the screeners were releasing based on risk. Instead, it is slightly *upward* sloping.

Figure 6 shows a somewhat more disturbing statistical relationship, by breaking this out by white defendants and black defendants. The arrest-rate/charge-rate relationship is flat for whites, but upward sloping for blacks. This is related to the result in Arnold et al. (2017), which reports the lower graph. Here, the x-axis is reversed such that the judges who are most severe are on the left. The graph shows that the “right” diagonal is found for white defendants, but there is a flat or slightly wrongly-sloped diagonal for black defendants.

Figure 4: Screeners Are Not Using Re-Arrest Risk

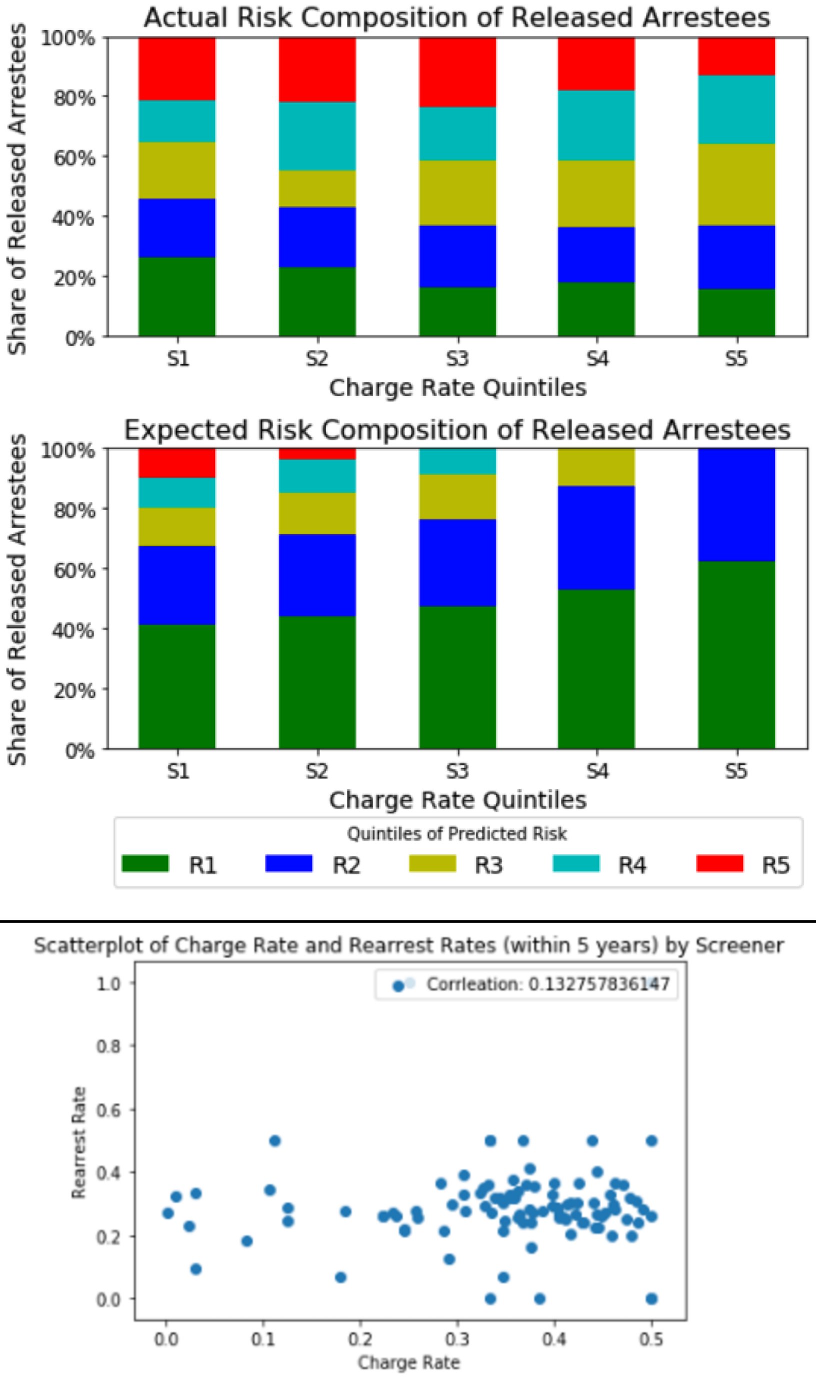
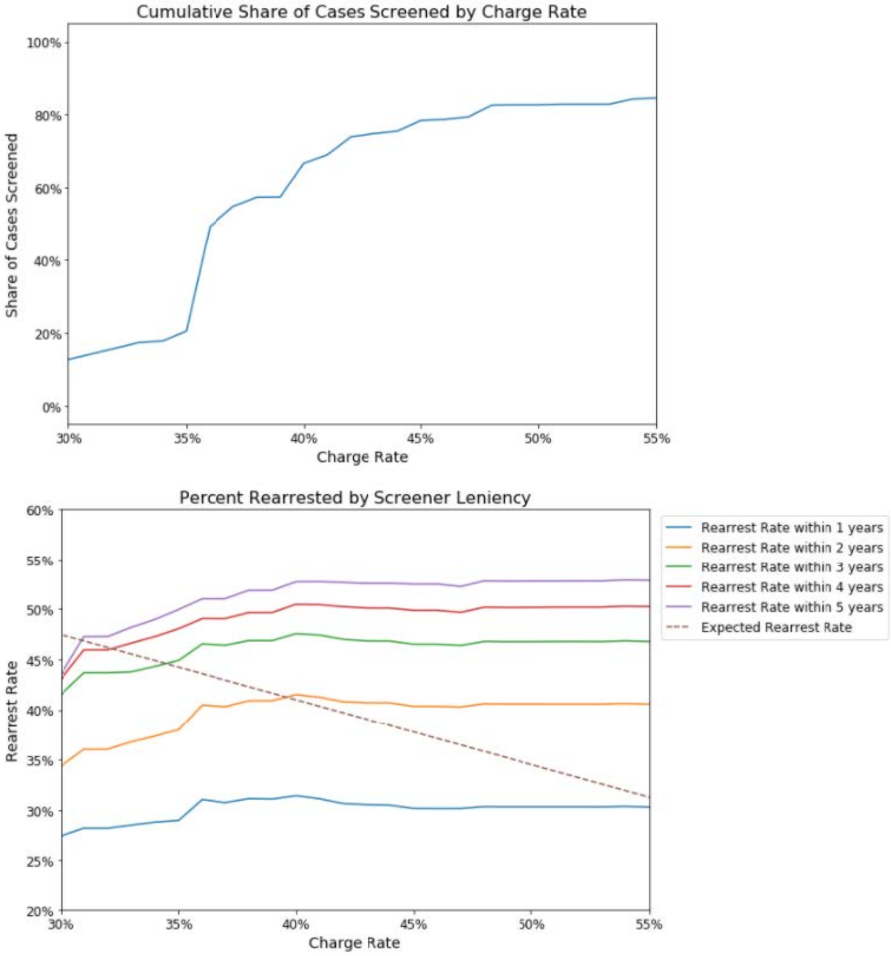


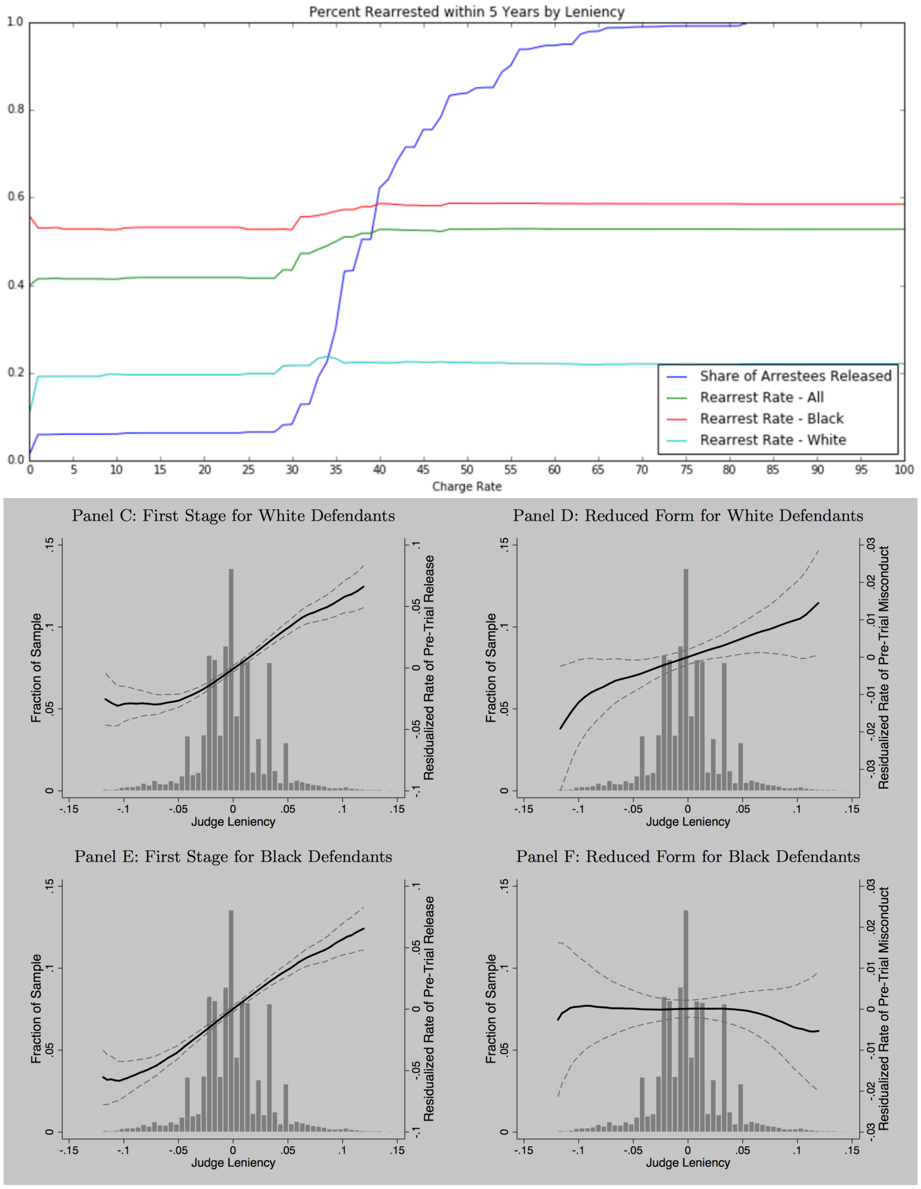
Figure 5: Screeners Are Not Using Re-Arrest Risk (cont.)



Cumulative Share of Cases Screened by Charge Rate

Percent Rearrested by Screener Leniency

Figure 6: Screeners Are Not Using Re-Arrest Risk (cont.)



Panel C: First Stage for White Defendants

Panel D: Reduced Form for White Defendants

Panel E: First Stage for Black Defendants

Panel F: Reduced Form for Black Defendants

4.2 Can the model help reduce recidivism?

Next we split screeners into two bins based on the percent of defendants they in fact charge, with a “lenient” group and a “strict” group. We compute the marginal number of defendants that would need to be charged for the “lenient" group of screeners to reach the same charge rate as the "strict" group of screeners. Using the fitted gradient boosted tree model, we estimate the probability of rearrest for the defendants released by the lenient screeners and rank them by risk.

We analyze the “marginal” defendant. Given a screener(s), we define the marginal defendant as the defendant with the highest predicted risk that was seen and released by that screener(s), based on estimated risk. If we arrive at a lower rearrest rate than the strict human screeners, then our model results in improvements in rearrest rates.

We arrive at a better outcome, in terms of rearrest rates, than the strict screeners. The potential improvements in rearrest rates are summarized in Figure 7. At a minimum, we reduce recidivism by 5 percentage points. At best, we improve by 9 percentage points.

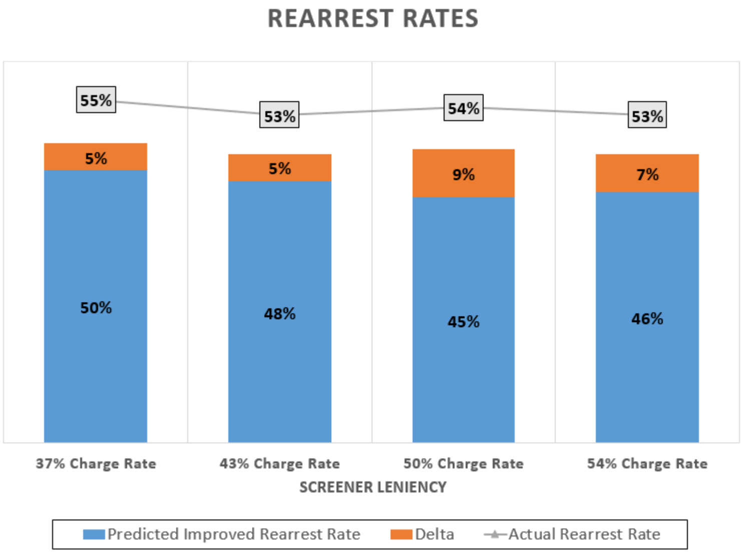
Our model did not result in greater racial disparities than human screeners. One reason may be the “wrong” slope for black defendants, in the previous section, as evidence of systematic racial bias in the system.

5 Discussion

5.1 Caveats

Unobserved variables like educational background, socioeconomic status, and a host of other factors are related to rearrest outcome. It is possible that different predictions (e.g., maximizing convictions, or sentence length, or minimizing time-to-trial) are driving the decisions. This does not, however, invalidate the comparison that we present. If a prosecutor’s office cares about rearrest rates, and one would assume that they all do, then the results of a successful rearrest algorithm should still be

Figure 7: Decreased Arrest Rates with Constant Charge Rate



relevant. Future analysis could run separate models for other charge and rearrest types.

5.2 Error Analysis and Time to Rearrest

We conducted error analysis to identify unusual features. First, we divided the observations according to whether they were correctly predicted and those that were incorrectly predicted and compared the distributions of every feature using box plots for continuous variables and bar plots for categorical variables. The result of the full comparison is in the appendix. We discuss one substantive finding here.

We observed that the incorrectly predicted observations had a similar distribution to correctly predicted observations for almost all features, with the exception of time to rearrest. The box plot below shows the distribution of time to rearrest for correctly predicted and incorrectly predicted observations in the positive class.

Figure 8: Distribution of Time to Rearrest for Correct and Incorrect Predictions

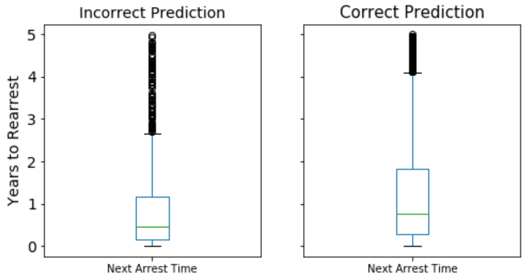


Table 5: Most Important Features

Column Name	Importance
AGE	0.281481
BAR_ADMIT_DAYS	0.167391
ARREST_TO_SCREEN	0.115987
SCREENING_DAYS	0.072958
CRIMINAL_FLAG	0.048115

To explain this pattern, we examined the top features by feature importance. In Table 5, we include the top 5:

The partial dependence plots of the top 5 features provide a clearer sense of the direction of the relationship.

We also examine the non-categorical features correlated with time to rearrest among the observations that were rearrested within five years. Below are the correlations between time to rearrest and the five non-binary, non-categorical features that have the strongest correlation with time to rearrest.

Figure 9: Partial Dependence Plots for Top Five Predictors

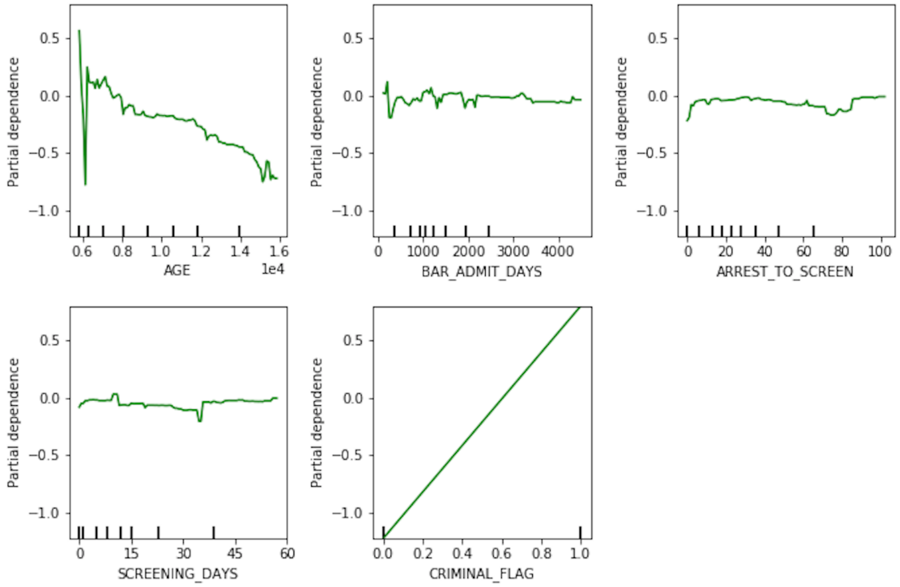


Table 6: Features with Highest Correlations with Time to Re-Arrest

Column Name	Correlation with Time to Rearrest
CRIMINAL_FLAG	0.229738
JUVENILE_FLAG	-0.184845
AGE	0.135949
INITIAL_DETENTION_FLAG	-0.037101
ARREST_TO_SCREEN	0.034602

Time to rearrest is positively correlated with criminal flag, and criminal flag is associated with a higher predicted probability of rearrest, which may explain in part why recall is higher among observations with a higher time to rearrest. Our results suggest that those arrested for criminal charges were rearrested later than those arrested for non-criminal charges. This finding may seem slightly counterintuitive, but further investigation may explain why the two are correlated.

5.3 Demographic Analysis

Perhaps the most immediately concerning factor that may be explored in other research is the importance of demographics in the model's success rate. In order to choose which features would optimize the model (including both categorical and non-categorical variables), we removed individual features one at a time and observed how the F-score changed. Sex, age, and race were consistently among the features whose removal had the highest negative impact on model performance. That criminality is associated with male gender and youth might not be especially surprising or concerning. There are likely unobserved variables at play which add bias to our model; educational background, socioeconomic status, and a host of other factors may affect the rearrest outcome but are only indirectly observable through the variables that we have and to which such class markers are related. Finding ways to include such unknowns would be an important way to improve the fairness of any resulting decision procedures.

5.4 Aligning the screener's and the algorithm's prediction

An important distinction between screeners and the algorithm is that while our model is classifying declined cases that result in rearrest to be 'wrong' decisions, a screener might not have been making the same prediction. It is possible that different predictions (e.g. whether or not a case can be tried and won) are actually driving the decisions at this node (Miller and Wright, 2008). To the extent these different predictions are correlated, the results of our algorithm might not differ that much. If a prosecutor's office cares about rearrest rates, the results of a successful rearrest algorithm should still be informative to them.

5.5 Potential Impact and Additional Steps

Given that the algorithm we developed predicted rearrest rates with relatively more accuracy than screeners, a potential implication would be to make it available to screeners as they consider declinations in real time. As pointed out by Kleinberg et al. in seminal research on the analysis of judicial bail decisions, such a model could provide decision makers with a risk score or flag for individual cases and serve as an aid, though not a replacement, for their judgment. Alternatively an algorithm could be used to rank entire populations of defendants for larger-scope recommendations. Large-scale ranking could be important when districts have to make assessments about the feasibility of caseloads and prison populations. While our model's success rate is promising to this end, an extensive amount of further research is required before a practical application could materialize.

One way to make the model more versatile would be to make it capable of identifying if lower charge rates can result in the same rearrest rate. This would allow policy makers to lessen the burden of cases with minimal impact on the level of rearrests.

Finally, we leave for subsequent research the relationship between screener traits and the characteristics of their declinations. Analyzing variation across screeners could identify whether or not biases exist based on screeners' political affiliation, age, experience, race, etc. One way of addressing this problem is to fit a model to predict screener declinations and to compare the feature most predictive in that model with the predictive features in the rearrest model.

5.6 Limitations

Incomplete data was a limitation in our study. We were unable to track the arrest registry beyond 1999 and unfortunately our requests for more recent data were not fruitful. Having access to the arrest records of the Orleans Parish jurisdiction would also have improved the model; it is highly likely that at least some of the defendants in the registry were rearrested in the future in another district but were identified by our model as correct decisions.

6 Conclusions

The decision to decline or pursue charges against a defendant has been identified as an potentially underemphasized point in the criminal justice process. Using eventual rearrest as an indicator of a successful declination decision, we created a model that outperformed human screeners on data from the NODA database. Applying this model to decisions at the declination node would have achieved lower rearrest rates between 5 and 9 percentage points, depending on the strictness level of the screeners we compared. Underlying biases in the model need to be addressed by including more explanatory variables, particularly with respect to demographic data of defendants and screeners. Using machine learning prediction algorithms to assess human decisions is a promising field of research in the court context and beyond, and further research such as this will hopefully have meaningful impacts on policy discussions.

References

Arnold, D., Dobbie, W., and Yang, C. S. (2017). Racial bias in bail decisions. Technical report, National Bureau of Economic Research. 4.1

Ash, E., Fagan, J., and Harris, A. (2017). Local public finance and discriminatory policing: Evidence from traffic stops in missouri. 2.1

Chen, D. L. (2017). How prosecutors exacerbate racial disparities: Screening gaps, race effects, and courtroom interactions. 1

Chen, D. L. and Phillippe, A. (2017). The long-run effects of criminal justice exposure on trust in the law and perceptions of legitimacy. Technical report. 2.1

Fagan, J. and Ash, E. (2017). New policing, new segregation? from ferguson to new york. *Georgetown Law Journal*. 1, 2.1

Fryer, R. G. (2016). An empirical analysis of racial differences in police use of force. Technical report, National Bureau of Economic Research. 2.1

Gopnik, A. (2017). How we misunderstand mass incarceration. *The New Yorker*, April 10, 2017. 1

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human decisions and machine predictions. Technical report, National Bureau of Economic Research. 1, 2.3, 4

Miller, M. L. and Wright, R. F. (2002). The screening/bargaining tradeoff. *Stanford Law Review*, 55(29):29–118. 1, 2

Miller, M. L. and Wright, R. F. (2008). The black box. *Iowa Law Review*, 94(1):125–196. 5.4

Mueller-Smith, M. (2014). The criminal and labor market impacts of incarceration: Identifying mechanisms and estimating household spillovers. Working paper, Columbia University. 2.1

Norris, S. (2017). Judicial errors: Evidence from refugee appealks. Technical report, Northwestern University. 2.1

Pfaff, J. F. (2016). Cheap on crime: Recession-era politics and the transformation of american punishment. 1

Rehavi, M. M. and Starr, S. B. (2014). Racial disparity in federal criminal sentences. *Journal of Political Economy*, 122(6):1320–1354. 1

7 Appendix

7.1 Data Merging and Cleaning

From the NODA database, we primarily used the following tables: Ada (prosecutor information), Areg (arrest register), Dfdn (defendant information), and Dsum (defendant summary). These tables were merged together using a combination of BOFI_NBR, DFDN_SEQ-_NBR, and ADA_CODE.

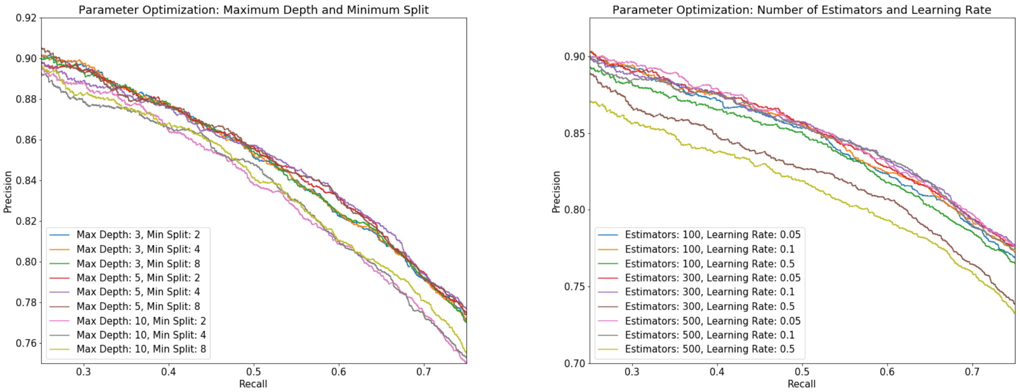
The dataset has undergone extensive cleaning and has documentation to explain which variables are reliable and which are less so. When possible, we imputed missing values based on the values of non-missing features (e.g., imputing the value of SCREENING_DISP_DATE from ARREST_DATE and SCREENING_days). For features where this was not possible, missing values were imputed at the mode. Using the codebook, we flagged invalid values of binary and categorical features and then transformed them using one-hot encoding.

We found that there were sometimes multiple arrests for a defendant on the same day. Some of these entries were duplicates while others took on different values for SCREENING_DISP_CODE. We flagged the multiple arrests so that there would only be one arrest per defendant per day, and if charges for that defendant were accepted for any of the arrests on that day, we set SCREENING_DISP_CODE to indicate that charges were accepted. Because our training sample excludes cases where the arrestee was charged, these re-coded cases were then excluded.

7.2 Model Validation Statistics

The precision-recall curves on the validation data are reported in Figure 10. The chart on the left shows varying tree depths and split thresholds for the optimal number of estimators and the optimal learning rate. The chart on the right shows varying estimators and learning rates for the optimal tree depths and split thresholds.

Figure 10: Parameter Optimization



7.3 Additional Summary Statistics

Here we report additional summary statistics for categorical variables.

CHARGE_CLASS	Freq.	Percent
1	1345	0.5
2	43597	16.17
3	116732	43.31
4	97363	36.12
5	27	0.01
6	6138	2.28
8	70	0.03
NA	4271	1.58

.

CHARGE_TYPE	Freq.	Percent
AR	153435	56.92
IF	113806	42.22
IN	1814	0.67
NA	488	0.18

.

PARTY	Freq.	Percent
D	82593	30.64
NA	51594	19.14
O	14136	5.24
R	121220	44.97

.

RACE	Freq.	Percent
A	1786	0.66
B	230672	85.58
H	1070	0.4
I	25	0.01
NA	2632	0.98
T	374	0.14
W	32984	12.24

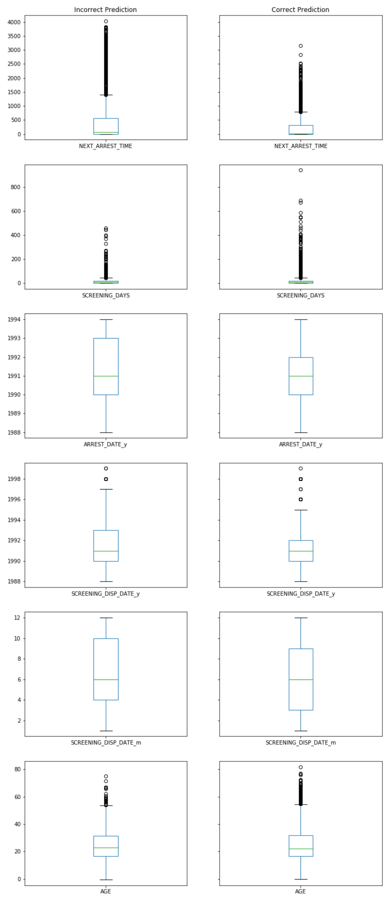
.

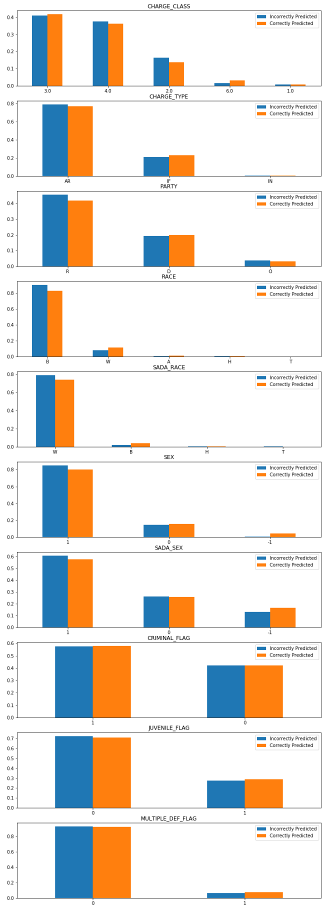
.

SADA_RACE	Freq.	Percent
B	17762	6.59
H	1408	0.52
NA	34380	12.75
T	4193	1.56
W	211800	78.58

7.4 Error Analysis

We presented substantive findings from the error analysis in the Discussion section. These plots demonstrate that time to rearrest was the only variable for which the distribution among correctly predicted observations noticeably differs from the distribution among incorrectly predicted observations.


Incorrect Prediction
Correct Prediction
NEXT_ARREST_TIME
NEXT_ARREST_TIME
SCREENING_DAYS
SCREENING_DAYS
ARREST_DATE_y
ARREST_DATE_y
SCREENING_DISP_DATE_y
SCREENING_DISP_DATE_y
SCREENING_DISP_DATE_m
SCREENING_DISP_DATE_m
AGE
AGE


CHARGE_CLASS
Incorrectly Predicted
Correctly Predicted
CHARGE_TYPE
AR
IF
IN
PARTY
R
D
O
RACE
B
W
A
H
T
SADA_RACE
W
B
H
T
SEX
SADA_SEX
CRIMINAL_FLAG
JUVENILE_FLAG
MULTIPLE_DEF_FLAG