NYU Center for Data Science

DS-GA 1003 Machine Learning and Computational Statistics

Can Gradient Boosted Trees Predict Rearrests More Accurately than Prosecutors?

Daniel Amaranto, Lisa Ren, Caroline Roper
NetIDs: da1933, tr1312, cer446

Spring 2017

Advisors ..... Daniel Chen, JD, PhD, and Elliot Ash, JD, PhD
Instructor ..... David Rosenberg, PhD

Abstract

We investigate how machine learning might bring clarity to a human decisions made during the criminal justice process. We created a model that predicts a defendant's risk of being rearrested after their charges are dropped. We used a database from the office of the Orleans Parish District Attorney that covers cases from 1988-1999. Applying strategies identified by past research that compared prediction models to judicial decision makers, our model selected higher risk individuals to prosecute than its human counterparts did. For a set charge rate, our model would reduce the rearrest rates between 5% and 9%. Developed further, such a 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.

Introduction

It has been claimed that the US criminal justice system incarcerates too many and generally does so unfairly. Experts have identified certain factors that have particularly strong effects on the system's purported inefficiencies and inequities, including the aggressiveness of prosecutors and the practice of plea bargaining. [1, 3]. It has been suggested that a reduced emphasis on plea bargaining and a heavier emphasis on the "screening" process—the stage when prosecutors decide whether or not to charge a defendant—could improve the fairness and efficiency of the system.[3] In our investigation we are not concerned with weighing in on the validity of such claims. We aim instead to augment the understanding of screening decisions made by prosecutors.

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. 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 [2] 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.

Data

Our data describe over a decade of arrests in a federal prosecutor's office from The Orleans Parish District Attorney's office. The current data set is from 1988 to 1999 and provides detailed information on approximately 430,000 charges and 280,000 cases (involving 145,000 defendants) filed or adjudicated during this timeframe. The data collected also contains detailed information regarding each individual offender, such as social security number and the corresponding prosecutor and judge.

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.

Data Cleaning

Data cleanliness is always a factor in analysis, but fortunately this dataset has undergone extensive cleaning already and has documentation to explain which variables are reliable and which are less so. When possible, we imputed missing values for continuous-valued features based on the values of other non-missing features (i.e., imputing the value of SCREENING_DISP_DATE from ARREST_DATE and SCREENING_days). For continuous-valued features where this was not possible, missing values were imputed as the mode. Using the codebook and lookup tables associated with the dataset, we flagged invalid values of binary and categorical features. Binary and categorical features were then transformed using one-hot encoding.

Our model only considers cases where the arrestee was not charged. 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. Those arrestees were then removed from the dataset fed into the model.

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.

Target Variable

To construct the target variable, we had had to identify arrests where the arrestee was rearrested within a certain number of years in the future. In order evaluate whether an arrestee was rearrested within a certain number of years, we truncated the data by that number of years from 1999, the last year for which we have arrest data. For example, if we wanted to create a target variable that indicates rearrest within 2 years, we would create that variable for defendants arrested between 1988 and 1997 so that we could conclusively determine whether an arrestee in 1997 was rearrested within two years. We created target variables for rearrests within one to five years.

Modeling

We followed the methodology and justification for the "Machine Learning Black Box" outlined in the Kleinberg paper to guide our modeling process.[2] Given a set of inputs, we train a model on the training data to predict the probability of rearrest, which we are using as a proxy for risk. 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 potential improvements in outcome, as measured by rearrest rates, if arrestees were charged based on their estimated risk.

Baseline Model

We constructed as decision tree of max depth 4 as our baseline model. We included the two features we believed would be most predictive: AGE and ARREST_CLASS. ARREST_CLASS indicates the severity of the charge on a scale from 1 to 7. The baseline model achieved 60% accuracy on the validation set and an F-score of 65%.

Model Optimization

Feature Selection

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. Repeat this process for 5 different target variables from rearrest within 1 year to rearrest within 5 years.
4. Evaluate each of the 220 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 removing Arrest Credit Code, keeping the rest of the features, and predicting rearrest within 5 years. The gradient boosted trees performed better than the random forest model on almost all feature combinations, including the best combination.

Parameter Optimization

After selecting rearrest within 5 years as the target variable and 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 ($\sqrt{features} = 9$, $log_2(features) = 6$)
3. max depth (8, 10, 12)
4. minimum samples split (2,4,8)

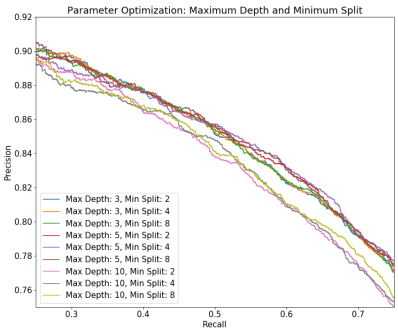
All of the random forest models use the gini index as the measure of the quality of split. The highest F-score for the random forest model was 0.7655, which is lower than the highest F-score for gradient boosted trees.

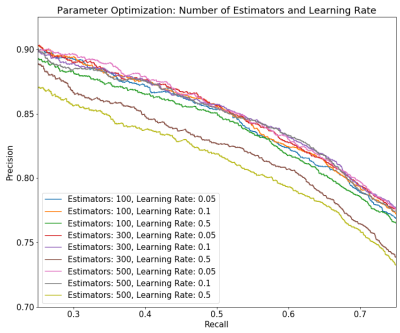
For the gradient boosted trees, we trained the model with the following values of hyperparameters for a total of 81 models:

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)

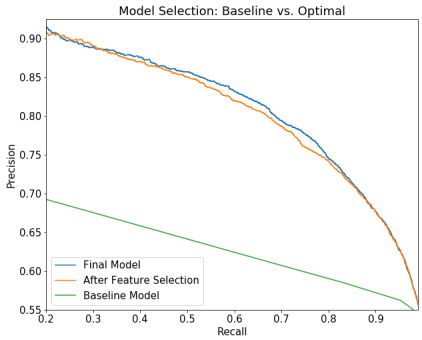
All the gradient-boosted tree models use deviance as the loss function to be optimized and the Friedman mean squared error as the measure of the quality of a split.

The resulting precision-recall curves on the validation data are below. 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.





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 validation F-score is 0.7703, which is .12 higher than the validation F-score of the baseline model. A comparison of the final model to the baseline shows the substantial improvement achieved.



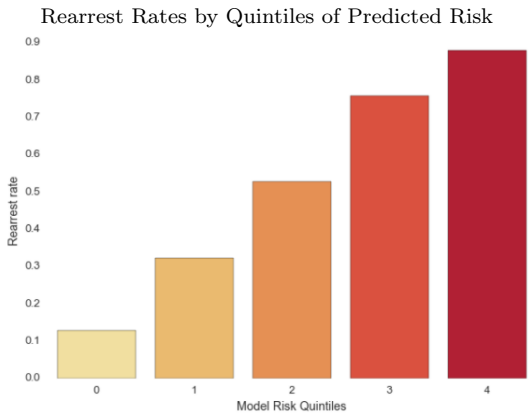
The final confusion matrix for the optimized model is below.

	Predicted No Rearrest	Predicted Rearrest
No Rearrest	3909	1387
Rearrest	1260	4438

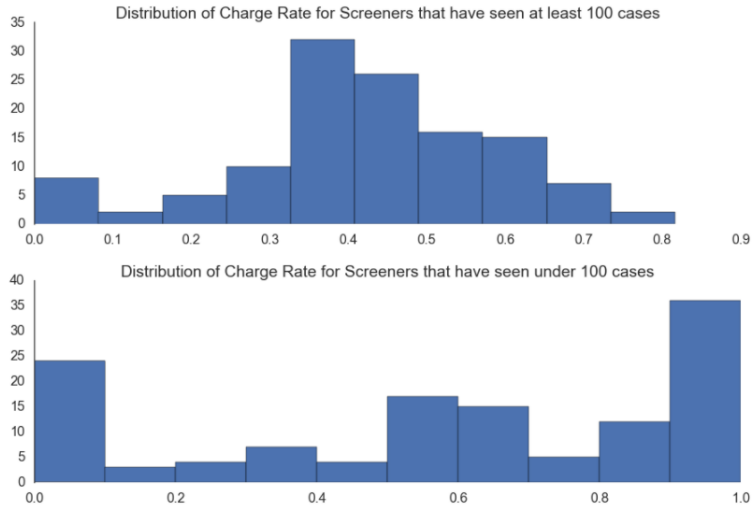
Note that all results are from the validation set. Because we plan to continue our research, we have not evaluated our model on the test set.

Results

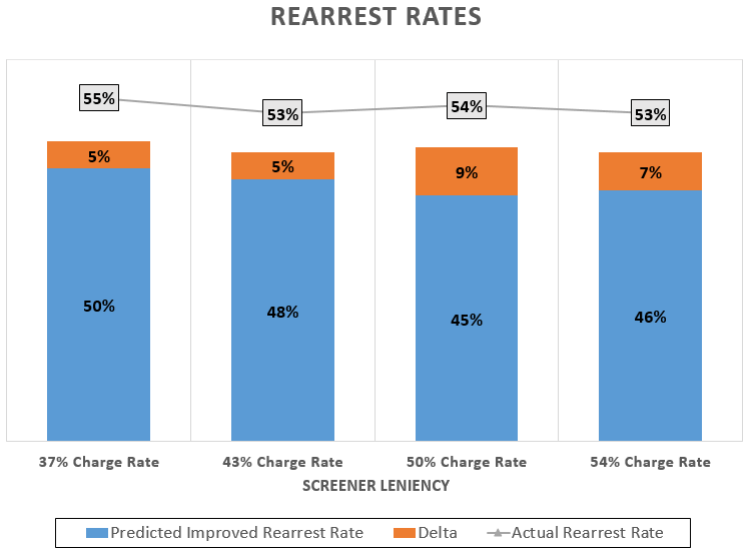
We assessed the actual risk of arrestees compared to the predicted risk returned by our model. Using the validation set, we grouped the arrestees into quintiles by their estimated risk and found that the predicted riskiest arrestees have higher rearrest rates. This shows that the arrestees that were released by a screener and predicted by our model to be risky were in fact risky.



We also assessed the performance of our model against human screeners using the reasoning derived in section 4.2 of the Kleinberg paper.[2] The NODA dataset provided sufficient variation in screener charge rates for us to analyze the marginal cases:



We group screeners into two bins based on the percent of arrestees that they charge. Then we calculate the marginal number of arrestees 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 arrestees released by the lenient screeners and rank them by risk. If we choose to charge the marginal number of arrestees based on the estimated risk from our model, we arrive at a better outcome, in terms of rearrest rates, than the strict screeners. The potential improvements in rearrest rates are summarized in the chart below.

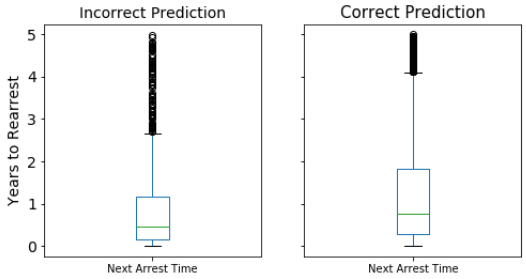


Discussion

Error Analysis and Time to Rearrest

We conducted error analysis to identify unusual features. 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.

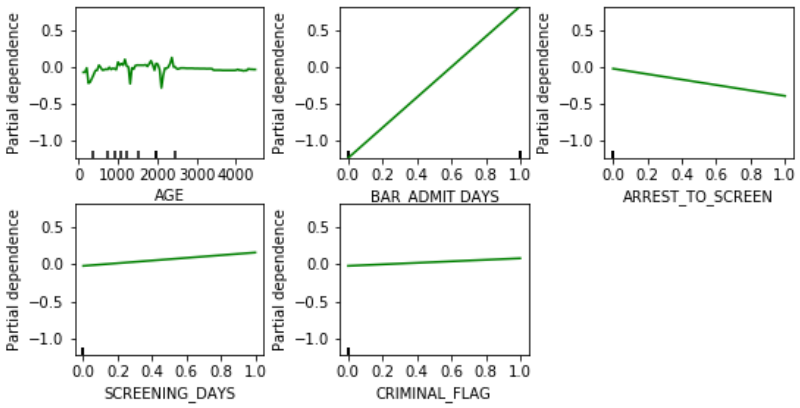
We observed that the incorrectly predicted observations had a similar distribution to correctly predicted observations for almost all features, with the exception of time to rearrest. The box plot below shows the distribution of time to rearrest for correctly predicted and incorrectly predicted observations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



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

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

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:

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. The correlation between time to rearrest and criminal flag is statistically significant according to the Pearson correlation coefficient's p-value. Our results suggest that those arrested for criminal charges who were rearrested 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.

Demographic Analysis

Perhaps the most immediately concerning factor that our research did not have time to address was 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, but if the model were

to exacerbate racial disparities by, for example, ascribing disproportionately less risk to white defendants, then it would need to be redesigned. It is all but certain that there are 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 to the model through the race variable to which such class markers are related. Finding ways to include such unknowns would be an important way to develop this analysis.

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.[4] 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 relevant to them.

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. [2] 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 were not able to analyze 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.

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.

Conclusion

The decision to decline or pursue charges against a defendant has been identified as an potentially underemphasized point in the criminal justice process [1, 3]. Using eventual rearrest as an indicator of a successful decliniation 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%, 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.

Acknowledgments

We wish to thank Dr. Daniel Chen and Dr. Elliot Ash for their extensive guidance throughout this project. The conception and execution of our analysis would not have been possible without their help.

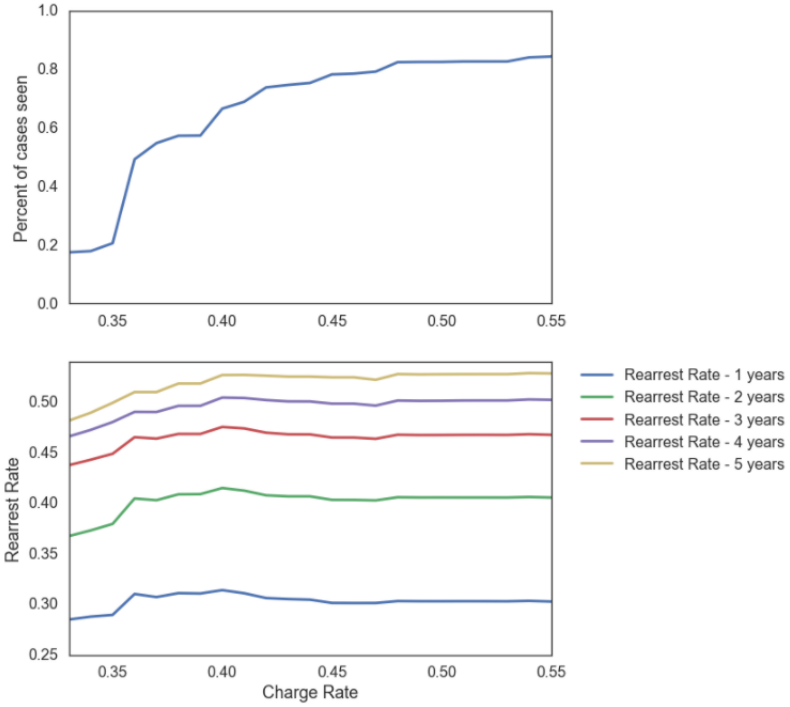
Appendix

Fall in rearrest rate as charge rate increases

To get a sense of the volume of decisions made by screeners of a given strictness level, we determined the cumulative percentage of cases seen as strictness increased. This revealed that the bulk of cases (roughly 60%) are seen by screeners with charge rates between 35% and 60%. Narrowing our focus to those screeners, we looked at the rearrest rates of their declinations across various time frames. As shown in the following plot, the rearrest rate reaches a peak at a charge rate of around 40% across several rearrest time frames. Then, remarkably, the rearrest rate falls before rising again and steadying. A possible explanation for this outcome is that cases are not randomly assigned to screeners, but rather they might be divided based on type of crime or some other factor. Screeners at charge levels from 0.4 to approximately 0.48 might be focusing on types of crime that are not associated with repeated criminal behavior.

Aside from the question of random assignment, this decrease in rearrest rate might be explained by separating these plots according to type of crime.

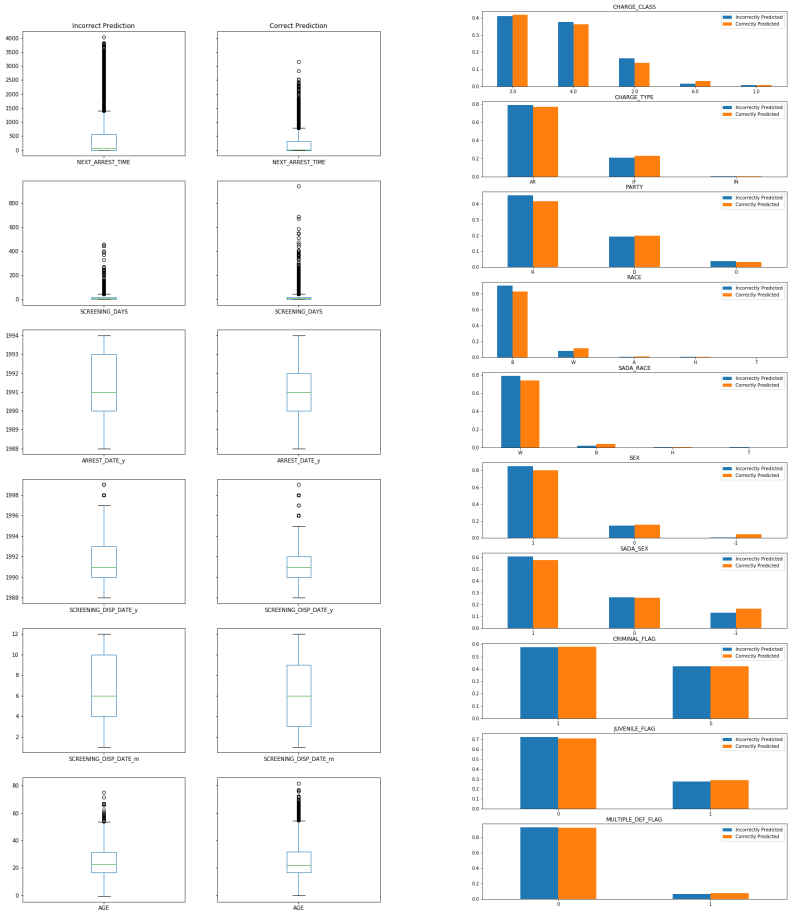
Rearrest rates and charge rates



The first graph shows the cumulative percentage of cases that are seen as more strict screeners are included. The bottom graph has the same x-axis, and shows the rearrest rate vs screener charge rate for 5 different rearrest time frames. Notice that around 60% of all cases are seen by screeners with charge rates between 35% to 60%.

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.



References

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

[2] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human decisions and machine predictions. National Bureau of Economic Research Working Paper Series.

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

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