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

Daniel Amaranto
Center for Data Science
New York University

Elliot Ash
Woodrow Wilson School of
Public and International Affairs
Princeton University

Daniel Chen
Institute for Advanced Study
Toulouse School of Economics
University of Toulouse

Lisa Ren
Center for Data Science
New York University

Caroline Roper
Center for Data Science
New York University

October 20, 2017

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 by applying strategies identified by past research that compared prediction models to judicial decision makers.

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 by the model would allow us to critique the way that screeners select defendants to charge.

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 information about the prosecutor, arrest register, and defendant.

To construct the target variable, we had to identify arrests where the arrestee was rearrested within a certain number of years in the future. In order to 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, removing 36,246 records (45%) from our training data.

Our approach starts by creating a model that takes as input instances of arrests in an arrest registry. Each arrest includes details related to both the charge and the arrested individual. Additionally, each arrest in the data has a field with the number of days, if it exists, until the individual has a subsequent arrest. Using data from which actual outcomes are known and a loss function that penalizes incorrect predictions, the model learns how to optimally predict the outcome if it receives a new input. Thus our model makes a prediction about whether each entry in a registry will or will not be rearrested within a set amount of time. The window of time during which we consider rearrest is variable. The model also generates a probability of rearrest for an instance; this probability, or risk score, is essential to formulating a comparison between the model and human decision makers. This prediction model was trained on a dataset of 80,217 arrestees. A validation set with 20,055 arrestees was used to evaluate the performance of different permutations of the model. A test set of 25,068 arrestees has still not been touched and awaits our completed model.

One of the simplest ways to predict whether someone would be rearrested is to use a decision tree based on age and the severity of the charge. Therefore we constructed a baseline model of maximum depth 4 using these two features. Given a parent node, we fit the model by minimizing the entropy of child nodes. Entropy is defined as

$$-(p_0 \log_2 p_0 + p_1 \log_2 p_1)$$

where p_0 is the proportion of the node comprised of arrestees who were not rearrested and p_1 is the proportion of the node comprised of arrestees were rearrested. The baseline model achieved 60% accuracy on the validation set and an F-score of 65%.

We refine our baseline model with Gradient Boosted Trees, a sequential ensemble model comprised of decision stumps. We used binomial deviance as our loss function:

$$\log(1 + e^{-2y\hat{y}})$$

where y is the true value of the target variable and $\hat{y}$ is the value predicted by the algorithm. Using this loss function allows us to interpret the score function as a probability.

Our best-performing final gradient boosted trees model used 500 tree estimators, a learning rate of 0.05, a maximum depth of 5, and a minimum samples split threshold of 4. The associated validation F-score is 0.7694, 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.

	Prediction: Not Rearrested	Prediction: Rearrested
Not Rearrested	3905	1391
Rearrested	1265	4433

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 (as in [2]) using the concept of implicit risk ranking. We do not directly observe how the screeners rank the risk of the arrestees that they see. 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. The NODA dataset provided sufficient variation in screener charge rates for us to assess the implicit risk ranking. We find that the actual risk distribution of risk among strict and lenient screeners differs from what we would expect to see if the screeners were releasing defendants based on their predicted risk.

Furthermore, we assess the potential improvement in outcome from using our model by analyzing the “marginal” defendant. Given a screener (or a group of screeners), we define the marginal defendant as the defendant with the highest predicted risk that was seen and released by that screener (or group of screeners).We group screeners into two bins based on the percent of arrestees that they charge. Then we calculate the additional 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 choose these “marginal” defendants based on estimated risk. We assess the outcome, in terms of rearrest rates, of charging these additional “marginal” defendants. If we arrive at a lower rearrest rate than the strict human screeners, then our model results in improvements in rearrest rates.

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.

In sum, the decision to decline or pursue charges against a defendant has been identified as a potentially under-emphasized 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. 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.

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.