Problem & Objective

> *"If computers could accurately predict which defendants were likely to commit new crimes, the criminal justice system could be fairer and more selective about who is incarcerated and for how long. The trick, of course, is to make sure the computer gets it right. If it's wrong in one direction, a dangerous criminal could go free. If it's wrong in another direction, it could result in someone unfairly receiving a harsher sentence or waiting longer for parole than is appropriate."* — ProPublica, "Machine Bias"

There are more than two million people incarcerated in the United States.[1] For the hundreds of thousands of people who enter prison each year, criminal risk assessment could be the difference between their spending years behind bars, and going free. Absent a tool to accurately gauge which defendants will re-offend, it is easy to err on the side of safety and imprison more people than we need to. On the other hand, a confident assessment that a defendant poses no risk to society would allow us to safely reduce sentences, place people under community supervision rather than incarceration, and potentially allocate resources for rehabilitation to those who need them most. Algorithmic risk assessment also stands to reduce the role of human bias in decisions about sentencing, bail, and parole. Even if it isn't perfect, it seems plausible that statistical models and/or artificial intelligence might be less swayed than a human judge by subjective and potentially discriminatory factors that are ultimately irrelevant to a defendant's risk of re-offending.

However, in practice, criminal risk assessment models have run into some serious problems. Most alarmingly, they have been found to be racially biased. A 2016 investigation by ProPublic found that COMPAS, a common risk assessment algorithm, failed *differently* for Black and white defendants: white defendants were twice as likely as Blacks to be labeled low-risk and then re-offend, while Blacks were twice as likely as whites to be labeled high-risk and then not reoffend.[2] It is especially difficult to justify the use of these models when their inner workings are opaque. COMPAS, for example, is owned by a private company, Northpointe, and the model behind the decisions it makes is proprietary.[3] Algorithms that are "black boxes" like this, whether it is because they are proprietary, or because they use such complicated techniques that they are not "interpretable" (e.g. deep learning and oodles of variables), make it difficult to justify or understand the decisions they are making.

Our starting assumption is that the ideal risk assessment algorithm would be not only accurate, but also simple and transparent enough to be interpretable. Any classification tool must make some assumptions about the individuals it classifies, but by removing these decisions from the black box, the public can deliberate and decide whether the features being considered are morally justifiable, and worrisome outcomes can easily be investigated. Our objective, therefore, is twofold: we want to develop a method of criminal risk assessment which has comparable accuracy to existing proprietary tools like COMPAS, but we also want that tool to be simple and interpretable (meaning that someone could look at the decision it made, and understand what factors were considered and how they were weighed).

Data & Behavior

As part of its investigation into COMPAS, ProPublica obtained a dataset of around 10,000 defendants from Broward County, Florida, which includes the variables considered by the model, the score given to the defendants, and a follow-up of whether they actually re-offended in the following years. This dataset is perfect for our purposes, because we can get learn a predictor on a chunk of the data, test it on the rest, and compare the results to the performance of the proprietary tool. In addition, we found a different dataset from New York (albeit with fewer features) which could might be useful to check whether our techniques generalize to a different place and a different dataset. The input to our predictor would be a list of features of a person, e.g. age, education, criminal history, etc. For example, we might have the input: {age: 31, race: "white", sex: "male", juvenile_felony: True}. The output could take one of two forms: (1) A binary classification $x \in \{0, 1\}$ where 0 means "will not re-offend" and 1 means "will re-offend"; or (2) A "risk score" (which could be continuous or discrete) which captures the idea that risk of re-offending is a spectrum, and gives a lower score to people less likely to re-offend, and a higher score to those more likely to re-offend. Because we want to experiment with different methods, we will likely try both classification and scoring, but will focus on classification in this proposal.

---

[1] Sawyer, Wendy, and Peter Wagner, "Mass Incarceration: The Whole Pie 2019," *Prison Policy Initiative*, 2019.
[2] Angwin et al., "Machine Bias," *ProPublica*, 9 March 2019.
[3] *Ibid.*

Previous Work

Besides the proprietary tools themselves, there has been a significant amount of criticism by journalists and academics, as well as attempts to develop alternatives. Many critics, including ProPublica (in an article entitled "Machine Bias"), point to the racial unfairness of these algorithms.[4] Even tools that don't explicitly consider race look at factors that are extremely racially skewed: for example, education level, previous convictions, and joblessness.[5] Another line of criticism, advanced forcefully by Cynthia Rudin, is that black box algorithms—where the 'black box' refers to either a trade secret or extreme complexity—should not be used to make high-stakes decisions. She rejects the widespread belief that more complex models are more accurate as a myth. Rather, she says, most of the time there are simple, interpretable alternatives, and that they should be preferred because we can actually *understand* them. This allows us to comprehend the outputs of the model, to see and rectify mistakes, and detect biases more easily.[6] Moreover, the complexity of black boxes makes them ripe for human error. For example, as Rudin observes: "COMPAS requires 130+ factors. If typographical errors by humans entering these data into a survey occur at a rate of 1%, then more than 1 out of every 2 surveys on average will have at least one typographical error."[7] Moreover, we would be unlikely to spot these errors: because the model is complex or opaque, it is harder to sanity check, because we literally do not know what it is doing.

Rudin also demonstrates that simple alternatives are realistic. Using an algorithm called "Certifiably Optimal Rule Lists" (or CORELS), she develops what is essentially a decision tree with three nodes based only on age, sex, and prior offenses. This algorithm is just as accurate as COMPAS, but unlike COMPAS, it is free and transparent, rather than a mysterious private software licensed and sold to the government.[8] Following Rudin, our aim in this project is to train and test models to predict recidivism that are *accurate*, *simple,* and *transparent*.

Baseline & Oracle

We developed an oracle and baseline for the classification task: labeling a defendant as "will re-offend" or "will not reoffend." For our baseline, we just predicted the most common label, which is "will not reoffend." This gave us a baseline accuracy of 54.5%. For our "oracle," we followed the suggestion of looking at the classification accuracy on training data. This is a good benchmark to shoot for because performance on test data is generally worse than performance on the training data to which the model was fit. Using this method with a logistic regression gave us an accuracy of 69.0%.

Potential Methods

The datasets we have gathered contain features such as age, race, marital status, and recommended supervision level during incarceration. We plan to implement a predictor for recidivism using several machine learning approaches, to the end of comparing their overall performance and gathering data on their racial bias. We have a couple potential ideas:

- *Decision Trees*: Since most relevant features have finite discrete values (e.g. male or female), we can create a decision tree that optimally "splits" these features to predict the likelihood of recidivism for a new data point. The decision tree will consist of internal nodes that interrogate a data point's features. Each internal node splits off into edges that lead to additional internal nodes, until there are no longer splits and a classification has been reached. One variation of this is the CORELS algorithm mentioned previously.
- *Logistic Regression*: Logistic regression is one of the most common and simple models for classification. Using the numerical features in our dataset, we will construct a logistic regression classifier and conduct a head-to-head comparison between our accuracy and that of COMPAS.
- After fitting these models, we will apply our implemented algorithms to a test set and report the accuracy of each algorithm, as well as their level of racial bias. More formally, this means comparing the false positive and false negative rate for different racial groups.

---

[4] *Ibid.*
[5] Casselman, Ben, and Dana Goldstein, "The New Science of Sentencing," *The Marshall Project*, 4 August 2015.
[6] Rudin, Cynthia. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” Nature Machine Intelligence 1 (2018): 206-215.
[7] *Ibid.*
[8] *Ibid.*

References

Angwin, Julia, et al. “Machine Bias.” ProPublica, 9 Mar. 2019, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Casselman, Ben, and Dana Goldstein. “The New Science of Sentencing.” The Marshall Project, The Marshall Project, 4 Aug. 2015, www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing.

Rudin, Cynthia. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” Nature Machine Intelligence 1 (2018): 206-215.

Sawyer, Wendy, and Peter Wagner. “Mass Incarceration: The Whole Pie 2019.” Mass Incarceration: The Whole Pie 2019 | Prison Policy Initiative, 2019, www.prisonpolicy.org/reports/pie2019.html.