Improving Accuracy and Reducing Bias in Criminal Risk Assessment

Benjamin Anderson
Stanford University
Stanford, CA
banders9@stanford.edu

Gaeun Kim
Stanford University
Stanford, CA
gaeunkim@stanford.edu

***Abstract* — Algorithmic criminal risk assessment has the potential to make the criminal justice system more fair, remove arbitrariness and human biases, and reduce incarceration by releasing people who we are confident will not re-offend. However, the current gold standard for criminal risk assessment (COMPAS) is severely lacking: an unaccountable and costly "black box" assessment which has also faced credible accusations of racial bias. Our goal in this project is to develop a simple, interpretable algorithm for criminal risk assessment. We show that a logistic regression, decision tree, and Bayesian network approaches all have comparable accuracy to the existing proprietary gold standard, while mitigating racial biases in misclassification.**

I. Introduction

There are more than two million people incarcerated in the United States [1], a number that has only increased over the years. 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. Even the most well-intentioned human judge is subject to subconscious biases along the lines of race, sex, appearance, and other factors. Algorithmic risk assessment could allow us to reduce the role of human bias in these high-stakes decisions about sentencing, bail, and parole, which determine the destiny of the millions of people who interact with the corrections system each year. These decisions are the difference between freedom and captivity, between being reunited with one's family and being alone for years, and between future potential and the near-permanent employment disadvantage that incarceration inevitably marks its victims with. These decisions also impact the basic safety of our society, as incorrectly releasing someone who goes on to re-offend puts other people's lives and property at risk. In short, the potential upside of accurate criminal risk assessment is enormous, and the social costs of mistakes—both false positives and false negatives—is similarly immense.

Despite the promise of criminal risk assessment systems, existing models—which have been widely adopted by various states and localities—have run into serious difficulties, which extend far beyond their accuracy. Most alarmingly, they have been found to be racially biased. A 2016 investigation by ProPublica found that COMPAS, a common risk assessment algorithm, failed *differently* for Black and white defendants: white defendants were twice as likely to be labelled high-risk and then not reoffend [2]. This is compounded by the fact that COMPAS is a *proprietary* system, meaning we cannot "look inside" to see what has gone wrong, as its inner workings of this algorithm are completely opaque. Our objective is to avoid these shortfalls by developing a method of criminal risk assessment which has comparable accuracy to existing proprietary tools like COMPAS, but is also simple and interpretable—meaning that someone could look at the decision it made, and understand how the factors considered affected the decision. We also want to avoid the racial disparities in classification that have plagued existing tools.

We leverage three models to tackle the problem: decision trees, logistic regression, and Bayesian networks. We find that each of these approaches has comparable accuracy to COMPAS. Moreover, each of them is *interpretable*: one can easily look at the model parameters and understand how various factors about an individual play a role in the decision. Finally, we analyze racial bias in misclassifications, and find that our approaches reduce racial bias relative to COMPAS.

II. Relevant Work

There has been a significant amount of criticism of proprietary risk assessment algorithms by journalists and academics, as well as attempts to develop alternatives. The focus of criticism has been the racial unfairness of these algorithms; even tools that don't explicitly consider race look at factors that are extremely racially skewed, such as education level and joblessness.

Another line of criticism advanced by Cynthia Rudin is that black box algorithms should not be used to make high-stakes decisions, because they prevent users from comprehending the outputs of the model, to see and rectify mistakes, and detect biases [3]. As a simpler alternative, she developed a decision tree with three nodes based only on age, sex, and prior offence. This free and transparent algorithm was just as accurate as COMPAS. Following Rudin, our aim in this project is to train and test models to predict recidivism that are accurate , simple, and transparent.

III. Problem and Approach

The dataset we used was published by ProPublica as part of its investigation into COMPAS. It consists of datapoints for 6,600 defendants from Broward County, Florida, including their COMPAS-assigned scores, ground truth values of whether or not they reoffended, and features related to age, criminal history, and demographics. From this dataset, we extracted binary features that can "split" the datapoints based on the following criteria: age under 25, age over 45, no prior crimes, one prior crime, fewer than five priors, fewer than ten priors, more than 20 priors, more than 30 priors, misdemeanour, African-American, Asian, Hispanic, Native American, White, other race, and female. We split the dataset into training (75%) test (20%), and validation (5%) sets.

To establish the baseline accuracy, which we later aimed to exceed in our classifiers, we created a model that simply predicts the most common label within the dataset for all individuals. The majority of individuals in the dataset we used did not reoffend, so "will not reoffend" was the universal prediction. Our baseline algorithm had a 54.5% accuracy.

For the oracle, which would establish a benchmark for ideal but not the most realistic algorithmic performance, we determined the accuracy of a logistic regression classifier applied to the training data. We decided this would be a reasonable way of 'overinforming' the oracle, since algorithm performance is typically worse for test data than for training data. The oracle achieved 69% accuracy. While not an algorithm we implemented ourselves, the COMPAS accuracy of 65% [4] was also a benchmark we intended to compare our algorithms against. Exceeding this accuracy would have significant implications for the criminal justice system; demonstrating that the complex, non-transparent algorithm is not the gold standard for recidivism prediction would motivate the transition to a fairer, interpretable model.

We implemented three predictors for recidivism using different machine learning approaches.

1. Logistic regression: One obvious choice for classification is a logistic regression, a model which learns coefficients for input features to predict the "log odds" that the output variable takes on one of its two possible values. Since we are trying to predict recidivism, logistic regression would find coefficients for the input features described before, in order to predict, given a new example, whether they would reoffend or not reoffend.
2. Certifiably Optimal Rule Lists: Since we converted our dataset to be a set of binary features with finite discrete values (e.g. male or female), we reasoned that it would be appropriate to implement a decision tree that optimally "splits" these features to predict the likelihood of recidivism for a new data point. We specifically used Certifiably Optimal Rule Lists (CORELs), which fits a decision tree (or "rule list") to a set of training data.
3. Bayesian network: Although Bayesian networks can be constructed based on prior (hopefully expert) intuitions about which factors influence which others, they can also be

learned directly from data via structure learning algorithms. Once the structure of the network and the conditional probability tables are inferred from data (which includes both the input features and the output, "reoffend" or "not reoffend"), a Bayesian network is a powerful tool for prediction. To use a Bayesian network for prediction, our query is the variable "will this person reoffend," and our evidence is the value of every other variable. Bayesian inference lets us compute the distribution over the query, and then we can predict whichever value has higher probability. In a Bayesian network, the probability of a variable depends on its parents, so we can see what factors influenced the decision by looking at the parents of the query variable.

We ran each classification algorithm with and without racial data in the feature set and compared accuracies. While race is a protected class and it is illegal to explicitly use racial information in risk assessment models, we were alarmed by the racial bias demonstrated in currently used algorithms and wanted to discern whether accounting for race in the feature set makes any difference in model accuracy. For further analysis, we also gathered the percentage of false positives and false negatives for each race, and directly compared these numbers against the COMPAS false classification rates published earlier by ProPublica.

IV. Results and Discussion

The three models we implemented all performed similarly when applied to the test set. CORELs, logistic regression, and the Bayesian network had accuracies of 66.7%, 66.6%, and 66.4%, respectively. These accuracies exceed the baseline accuracy of 54.5%, and notably exceed, in all three cases, the COMPAS accuracy of 65%. Given the simplicity and interpretability of these algorithms, and our parsimonious feature set, it is surprising that they perform as well as a complex proprietary algorithm that considers more than 130 variables (Figure 1).

Accuracy of algorithms in training and test set

	CORELs	Log regression	Bayesian net
Training	0.656	0.654	0.654
Test	0.667	0.665	0.665

Figure 1: All three implementations beat the baseline accuracy on the test set.

We then removed all race-related data from the feature set, and observed whether our model accuracies significantly changed. The removal of racial data had a negligible impact on accuracy. CORELs,

logistic regression, and the Bayesian network were 66.7%, 66.5%, and 66.5% accurate—there was no more than a 0.2% decrease in accuracy when race was omitted from consideration (Figure 2).

From these basic tests of accuracy, we concluded that our interpretable models match COMPAS in terms of accuracy. We also confirmed that using racial data as features has almost no effect on algorithmic accuracy, quantitatively justifying the exclusion of this data in future data collection.

Accuracy of algorithms with and without racial data

	CORELs	Log regression	Bayesian net
Racial data	0.667	0.666	0.664
No racial data	0.667	0.665	0.665

Figure 2: Accuracy of algorithms does not depend on the inclusion of individuals' race information.

Following the implementation of the Bayesian network model, we visualised the network and examined the parent-child relationships between variables (Figure 3). The Markov blanket of the variable of predictive interest (whether an individual reoffended in the two years post-conviction) extended to 'fewer than five priors' and 'greater than thirty priors', indicating that the only factor conditionally dependent on the likelihood of reoffence is criminal history. Notably, this Markov blanket was exclusive of any racial or demographic variables. The network was divided into two components. All racial variables other than 'Race: other' were connected in a component completely disparate from the two-year recidivism variable.

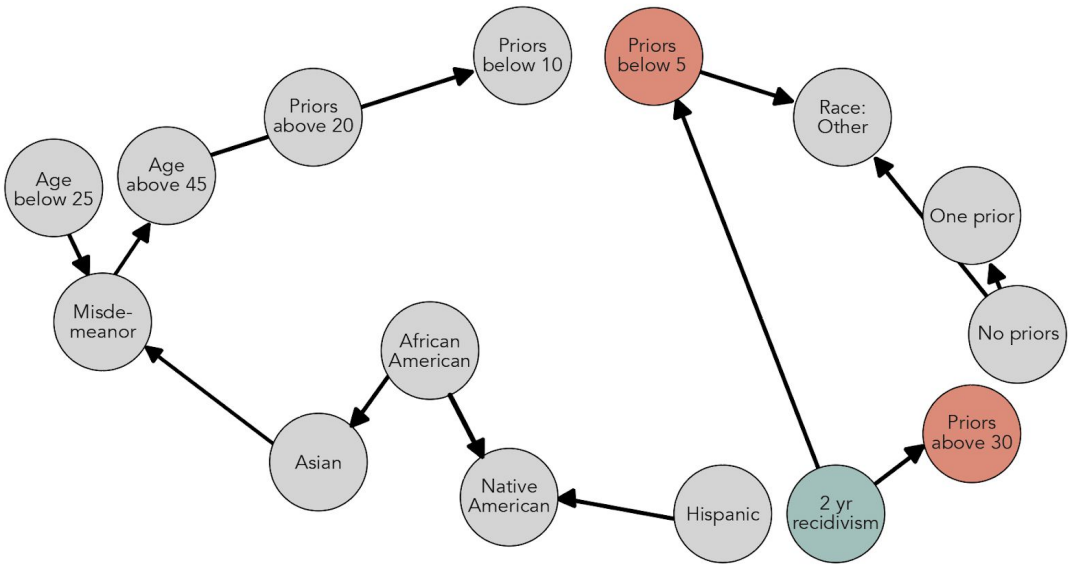
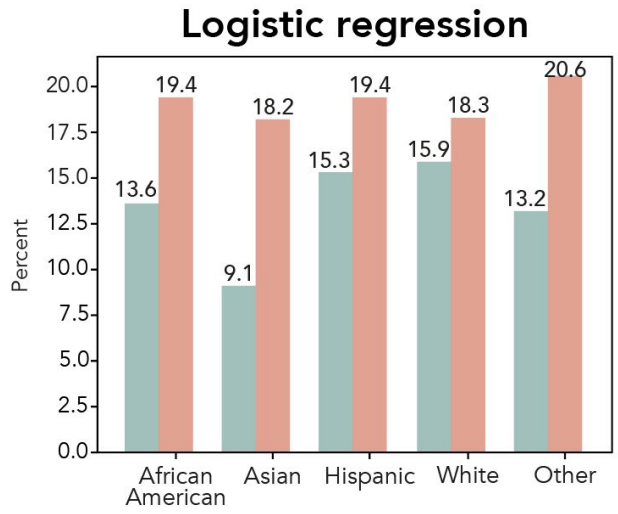
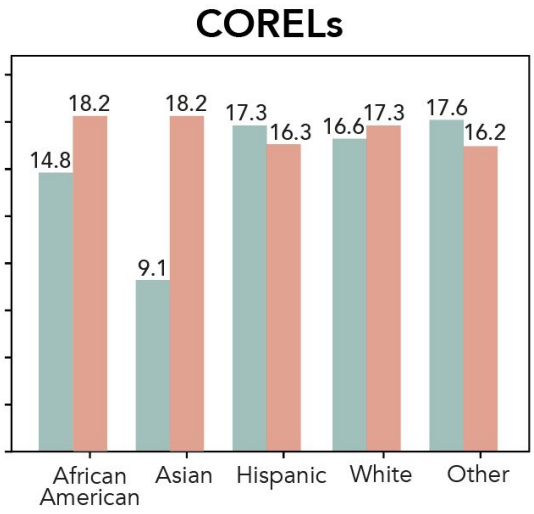
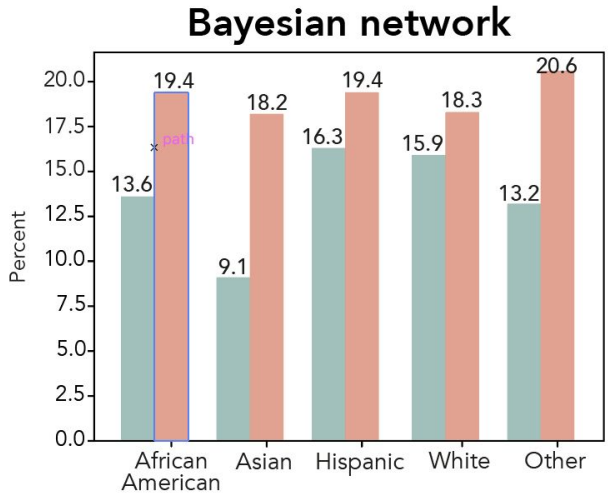


Figure 3: Visualization of Bayesian network for the dataset including racial variables.

We calculated the algorithms' percentage of false positives (labelled high risk, did not reoffend) and false negatives (labelled low risk, did reoffend) for each racial group, as a way of uncovering any racial bias built into the classifiers. For instance, if the false positive rate were much higher for defendants of minority races than for white defendants, using the algorithm in the real world would result in minorities being unfairly sentenced and denied parole. The logistic regression model yielded a 13.6% false positive rate for African American defendants, and 15.9% for white defendants. These rates were 14.8% and 16.6% for CORELs, and 13.6% and 15.9% for the Bayesian network. In all three models, non-white defendants were not at all disadvantaged in being labelled high risk and not re-offending. In fact, individuals' race appeared to be uncorrelated with false positive and false negative rates altogether (Figure 4). This was a significant improvement from the COMPAS algorithm, which has been shown to be heavily biased against African American defendants, giving them a 44.9% false positive prediction as opposed to 23.5% for white defendants.







COMPAS false classifications

	White	African American
Labeled high risk, didn't re-offend	23.5%	44.9%
Labeled low risk, did re-offend	47.7%	28.0%

Figure 4: Analysis of how race is related to errors in our classifiers, compared to COMPAS false classification rates.

V. Conclusion

We set out to develop models for criminal risk assessment whose accuracies are on par with the accuracy of the gold standard commercial algorithm. At the same time, we wanted to determine whether racial biases evident in currently used proprietary algorithms are the unavoidable consequence of biases in prosecution, or the result of the unnecessary complexity of current classification models. Upon implementing three models—logistic regression, CORELs, and Bayesian network—we found that these simple classifiers surpassed the accuracy of COMPAS and were not made any less accurate by exclusion of racial variables. Direct comparison of our algorithms' false positive and false negative rates with those of COMPAS showed that we made significant improvements in the racial parity of predictions.

References

[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] 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.
[4] Yong, Ed. "A Popular Algorithm Is No Better at Predicting Crimes Than Random People." The Atlantic. 17 January 2018. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/.

Data & Code

https://drive.google.com/open?id=1MqCw_8HOEGlVgd-Rl9pJ0ZJ4Yv_DnnZG