Improving Accuracy and Reducing Racial Bias in Criminal Risk Assessment

Benjamin Anderson (banders9), Gaeun Kim (gaeunkim)
CS 221 (Aut 2019) Final Project

Motivation

"If computers could accurately predict which defendants were likely to commit new crimes, the criminal justice system could be fairer ... about who is incarcerated and for how long." – ProPublica, "Machine Bias"

- 2M+ incarcerated in the US; could imprison fewer people with accurate risk assessment
- Current gold standard: COMPAS
  1. Black box algorithm (lack of public accountability)
  2. Racially biased

Problem Definition

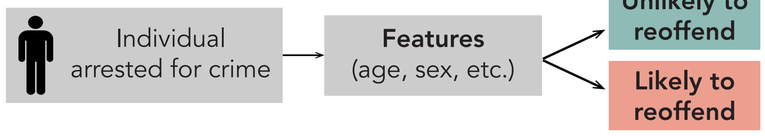


Figure 1: Schematic of a generic reflex model appropriate for our project goals.

Can we develop and compare algorithms for criminal risk assessment that a) have comparable accuracy to existing propietary tools and b) are interpretable and racially unbiased?

Baseline: Predicting the most common label

Ground truth: {0, 0, 1, 0, 1, 0, 0, 0, 1}
Predictions: {0, 0, 0, 0, 0, 0, 0, 0, 0}

54.5% accuracy

Oracle: Accuracy of logistic regression on training data

Good benchmark since performance is typically worse for test data than for training data

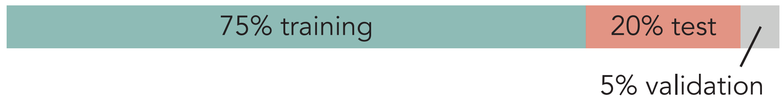
69% accuracy (COMPAS accuracy: 65%)

Methodology

Extract binary features from ProPublica dataset

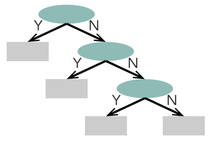
Age-related	[under 25, over 45]
Crime-related	[no priors, 1 prior, <5 priors, <10 priors, >20 priors, >30 priors, misdemeanor]
Demographics	[African-American, Asian, Hispanic, Native American, White, other race, female]

Split the data (n = 6600)



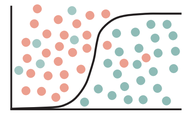
Decision tree

Minimize (fraction of samples that the rule list misclassifies) + (factor ~ length of rule list)



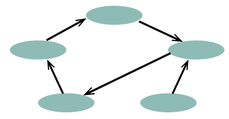
Logistic regression

Learn coefficients for input features to predict the "log odds" that the output is 'will reoffend' or 'will not reoffend'



Bayesian network

Compute distribution over the query "will this person reoffend?" and find most probable values of other variables



Results: Accuracy

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

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

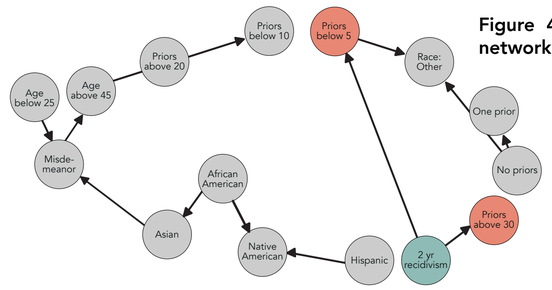


Figure 4: Visualization of Bayesian network for the race-blind dataset.

We can observe that the Markov blanket of the 2 year recidivism variable (highlighted in red) is exclusive of any racial variables. In fact, the component of the network which includes racial variables is disconnected from recidivism altogether.

Results: Racial Bias

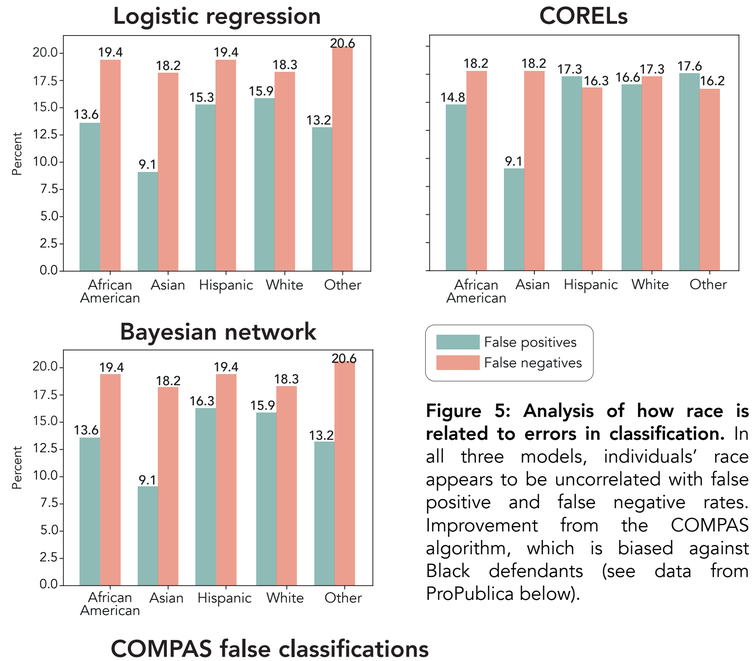


Figure 5: Analysis of how race is related to errors in classification. In all three models, individuals' race appears to be uncorrelated with false positive and false negative rates. Improvement from the COMPAS algorithm, which is biased against Black defendants (see data from ProPublica below).

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%

Note significantly higher false positive classification rate for African American defendants than for white defendants.

References

1. Angwin, Julia, et al. "Machine Bias." ProPublica, 9 Mar. 2019.
2. 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.
3. Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." arXiv preprint arXiv:1609.05807 (2016).