Handout Recap: ProPublica's "Machine Bias" Report

Kevin Klein, Yorick Scheffler

May 4, 2023

1 Introduction

In 2016, investigative journalism nonprofit newsroom ProPublica published a report titled "Machine Bias - There's software used across the country to predict future criminals. And it's biased against blacks.". The report gained a lot of media attention and sparked a debate about the usage of algorithms in the criminal justice system. The report sheds a light on how those systems can perpetuate already existing bias and racial injustice in the system. ProPublica focuses in their report on one proprietary algorithm called "Correctional Offender Management Profiling for Alternative Sanctions", in short "COMPAS". This solution is developed by Northpointe and in use widely across different states and jurisdictions. The algorithm classifies criminals based on 137 questions across a broad range of topics, where race is no explicit question. It calculates a risk score for them from low risk=1 to high risk=10.

ProPublica's findings are based on a statistical study with 7,000 individuals who were arrested in Broward County, Florida between 2013 and 2014. In the study they found out that black defendants were scored with a higher risk than white people with a similar or worse criminal history (see figure 1). Likewise the algorithm was more likely to flag white people at a lower risk than they actually were, while the COMPAS algorithm achieves an overall accuracy of around 65%. The findings persisted even when ProPublica controlled for other factors like age, gender, and prior convictions.

The report was the starting point of a national debate about the usage of such algorithms and the risk of reinforcing already existing bias and injustice. Northpointe reacted to the publication, rejecting the accusations of unfair risk calculation across ethnic groups and claiming that their product is equally accurate for black and white defendants. They criticized the statistical methods used in ProPublica's article as being insufficient to prove unfair treatment, because they fail to take different base rates of recidivism into account. Furthermore, Northpointe examined the same dataset and applied different statistical methods which they claimed were more suitable to find evidence of bias. Unsurprisingly, the results showed no signs of bias in the software's risk assessment for blacks and whites.

The debate between Northpointe and ProPublica about the accuracy of their statistical analysis attracted the attention of researchers across the USA, which had some significant findings that were a stepstone for examining algorithms in regards to coming closer to a "true fairness" and showing

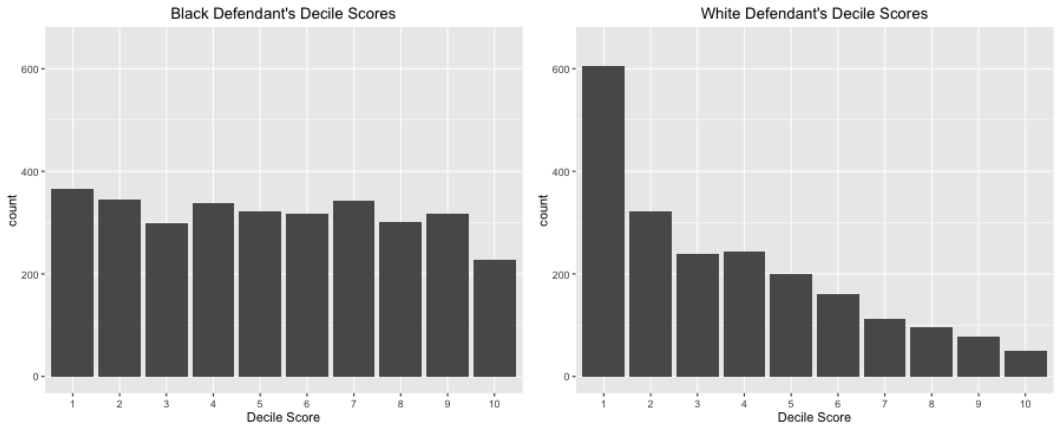

Figure 1: Unequal distribution of risk scores for black and for white defendants, as reported by [ALKM16]

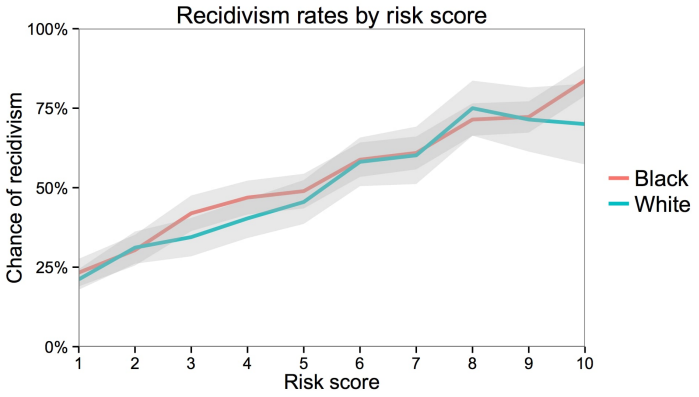

Figure 2: Chance of recidivism per risk score, as illustrated by [CDPFG16]

that the findings aren't as black and white as they appeared. Over months, various research groups examined the case and evaluated different statistical methods as well as their possible interpretations. As it turned out, ProPublica and Northpointe were using different types of fairness as a basis to prove or disprove bias in COMPAS' risk scores, hence disagreeing with the analyses of each other.

ProPublica addressed disparities on the level of group fairness, stating that unequal distribution of risk scores among people of different ethnic groups and different error rates for non-reoffenders are clear indicators for bias. On the other side, Northpointe laid their focus on evaluating the accuracy of their software, hence targeting individual fairness. Indeed, when COMPAS calculated a risk score for an individual, the probability of recidivism in a real-world dataset was equal across races (see figure 2), ages and other common discriminatory factors.

Researchers even showed that these two different perspectives of fairness on the results of a risk assessment system like COMPAS are unavoidable, meaning that when optimizing a software to comply with one type of fairness, it will always violate the other ([CDPFG16]). The problem is that the underlying data is heavily influenced by cultural stereotypes, wealth distribution, level of education and inequalities in the police and juridical system of the United States. Aiming for maximized accuracy, as Northpointe did, will incorporate these flaws into a risk assessment software. However, targeting group fairness will reduce its accuracy, which might make it less interesting for lawmakers and legal actors to employ the tool.

Since 2016, this story has been used many times as a case study for topics such as assisted and automated decision making, risk assessment and predictive policing. It fueled controversies about fairness and bias in software and attracted the attention from legislators around the world. The debate shed a light on the difference of group fairness and individual fairness. This has revealed the inherent trade-offs between optimizing for accuracy and avoiding bias, highlighting the need for considering the potential cultural and systemic biases inherent in the data.

References

[AL16] Julia Angwin and Jeff Larson. Bias in criminal risk scores is mathematically inevitable, researchers say, Dec 2016.

[ALKM16] Julia Angwin, Jeff Larson, Lauren Kirchner, and Surya Mattu. Machine bias, May 2016.

[CDPFG16] Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear., Oct 2016.

[FBL16] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

[LAKM16] Jeff Larson, Julia Angwin, Lauren Kirchner, and Surya Mattu. How we analyzed the compas recidivism algorithm, May 2016.