Amro Ashmeik

November 12, 2019

STAT 370 Human Rights Statistics

Professor Bruce D. Spencer

A Literature Review of Algorithmic Bias

Introduction

Recently, Apple came under fire after David Heinemeier Hansson, Danish programmer and the creator of the popular Ruby on Rails web development framework, tweeted that his wife's application for a credit line increase on her Apple Card was declined, despite her having a better credit score and other factors that would contribute in her favor. Steve Wozniak, an Apple co-founder, tweeted about a similar occurrence with him and his wife. Hansson said that when his wife contacted Apple about the issue, they told her that they were not authorized to discuss the credit assessment process. New York State regulators announced that they would investigate the credit worthiness algorithm used by Apple Card for potential discriminatory treatment, regardless of whether it is intentional or not. Discrimination against women in assessing credit worthiness is just one example of how machine learning and the use of algorithmic decision making has reinforced societal biases.

Machine learning is a tool that enables the automation of statistical model building. It is used in a wide variety of industries and government functions and continues to expand in its applications as the world becomes more data-driven. The focus of this paper will primarily be on the state of machine learning tools in the contexts where the outcome of the machine learning model could have disparate effects on the groups of people it is applied to. Examples of these contexts include the hiring process, automated risk assessment by U.S. judges to determine bail and sentencing limits, and financial services discrimination such as in the case of the Apple Card story or mortgage lending a.k.a. “algorithmic lending.”

A dangerous assumption made with regards to these models is the assumption of “neutrality”—that unlike a human decision maker, a computer won’t be biased. It is a dangerous assumption to make because often the data used in generating statistical models is biased in itself. Data that is human generated will carry human biases that will translate to the models that are developed. Biased machine learning models risk to deepen socio and economic disparities and so action must be taken to mitigate the effects of algorithmic bias. This paper is a literature review of algorithmic bias, addressing the causes, the issues that it raises, and ways bias can be statistically controlled for and legally held accountable.

Cases of Algorithmic Bias

Predictive Policing

An automated tool being increasingly used in the U.S. is predictive policing systems. The RAND (Research and Development) Corporation defines predictive policing as “the application of analytical techniques – particularly quantitative techniques – to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions”. Research has shown that police databases lack in quality, carrying biases that would be reflected in models trained on those databases. The Royal Statistical Society (RSS) conducted a study in Oakland, CA to assess how just how biased police databases are and if predictive policing systems would encourage policing of neighborhoods that are already overrepresented in historical police data, in which the software will have failed to adjust for the apparent biases in the data.

To test the claims that police databases were biased, the RSS combined a synthetic population of Oakland, CA with data from the 2011 National Survey on Drug Use and Health (NSDUH). A synthetic population is a demographically accurate individual-level representation of a real population such that the demographic characteristics of the population matches data from US Census Bureau as closely as possible. The combined dataset acted as a "ground truth" for estimates of drug use in Oakland and allowed researchers to create a map that showed the distribution of drug use across the city. The validity of using this dataset as a ground truth comes from the fact that the US Bureau of Justice Statistics – the government body responsible for compiling and analyzing criminal justice data – has used data from the NSDUH as a more representative measure of drug use than police reports. In addition, data from the NSDUH is a more statistically representative sample of illicit drug use in the population than police records. The RSS found that arrests related to drug use were concentrated in particular parts of the city that had large populations of low-income individuals and minorities. Meanwhile, a map of drug use from the ground truth dataset showed a much more spread out distribution of drug use. The RSS states that "These neighborhoods experience about 200 times more drug-related arrests than areas outside of these clusters." The conclusion on the data was that "police data appear to disproportionately represent crimes committed in areas with higher populations of non-white and low-income residents." Figure 1 below shows the disparities in the distributions of actual drug use in comparison to arrests related to drug use.

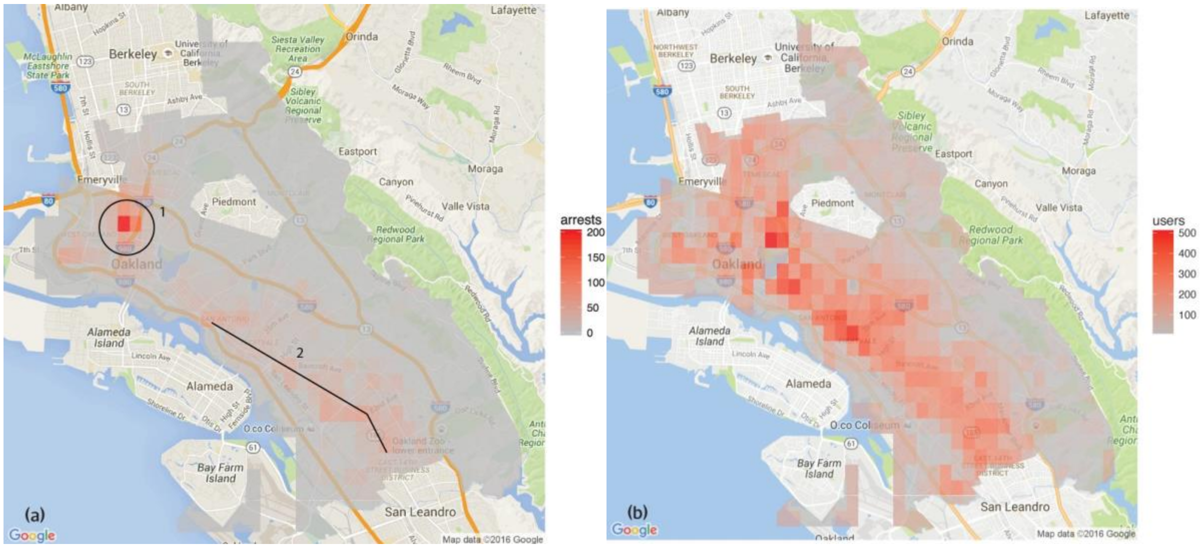

Figure 1

(a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

A question/criticism I do have is that the RSS does not explain what data they pulled from police databases when compiling their dataset of drug arrests. It could be the case that the drug arrests they pulled include violent drug offenses. In that case, the type of “drug use” occurring would be different between the datasets being compared. Only aggregating non-violent drug arrests would be a more accurate representation of portraying disparities in police enforcement of illicit drug use.

After concluding that disparities exist in how the police enforced drug laws, the RSS was interested in how this biased training data would affect predictive policing systems. The hypothesis is that since the particular neighborhoods and ethnic groups were overpoliced, the algorithms trained on police data would direct officers to police those same neighborhoods leading to a positive feedback loop. The RSS used a publicly released predictive policing algorithm developed by PredPol, one of the largest vendors of such software, to test their hypothesis. They came to the conclusion that the software lead officers to disproportionately police low-income neighborhoods and communities of color.

The results presented thus far relied on the assumption that areas that saw increased policing did not see an increase in crime/arrests. To illustrate the feedback loop hypothesized earlier, a simulation was performed where the number of crimes was artificially increased by 20% wherever the PredPol software assigned additional policing. The simulated crimes were added to the training data for the PredPol algorithm which then relearned and adjusted its targets for extra policing in the following days. Police presence should not be increased in areas that saw increased arrests due to targeted policing, otherwise you would see a positive feedback loop. Nevertheless, the simulations conducted by the RSS showed that the additional arrests that resulted from targeted policing recommended by the algorithm led to a positive feedback loop where the algorithm would further increase police targeting where arrests increased. Figure 2 below illustrates the result of the simulations.

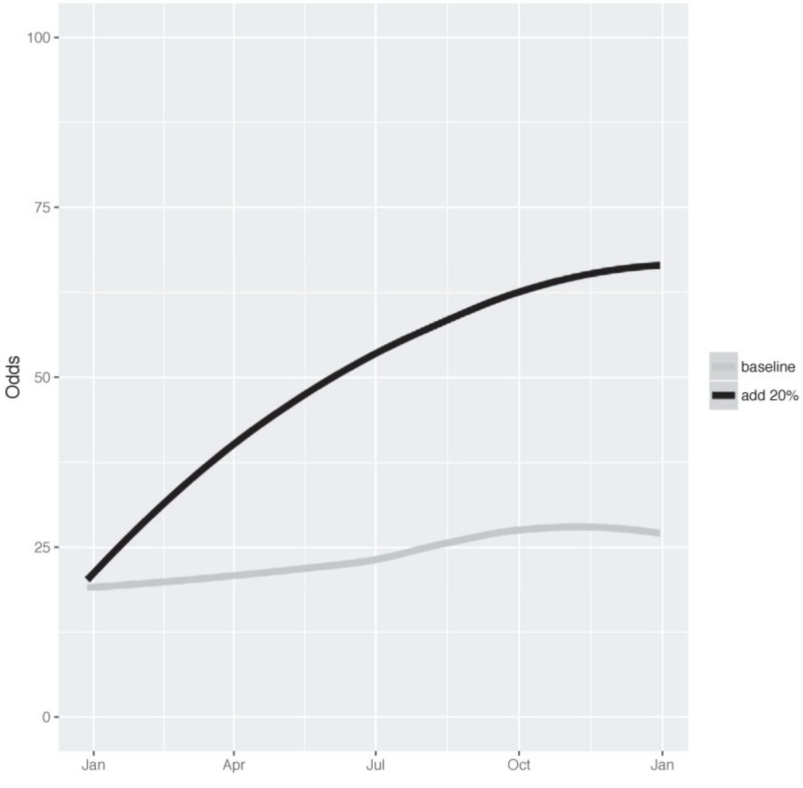

Figure 2
Predicted odds of crime in locations targeted by PredPol algorithm, relative to nontargeted locations. ‘Baseline’ is original Oakland police data. ‘Add 20%’ simulates the effect of additional crimes being observed in targeted locations

AI in the Hiring Process

Hiring is another process where automation has been lauded as a method for objectively and efficiently evaluating candidates. As stated before, the assumption is made that because a machine, not a human, is making or aiding in the decision-making process of hiring, it is non-discriminatory. Nevertheless, and unsurprisingly, algorithmic bias has also been shown in the hiring process.

One example is in the case of Amazon. Despite being a technological leader in AI, it suffered from biases in their hiring algorithms. Amazon’s AI hiring software did not rate candidates for technical positions in a gender-neutral way. The AI was trained on resumes over a 10-year period, most of which came from men reflecting male dominance in the tech industry. As a result, Amazon’s system taught itself to prefer male candidates, penalizing resumes that included words like “women’s”. In response, Amazon altered their model to treat such words as neutral, but the model could still build other associations that biases against women. Ultimately, Amazon killed the project and stated that ““This was never used by Amazon recruiters to evaluate candidates.”

Another instance of hiring bias occurred when Xerox hired data scientists to assist with the hiring process for their call centers. The models the data scientists developed revealed that the most important factor to employment retention was the employee’s distance from work. The data scientists quickly found that an employee’s distance from work was strong correlated with race and thus, using it as a factor would be discriminatory.

The trend seen here with using algorithms in the hiring process is that models learn patterns in the data that are unintentionally discriminatory and it's usually as a result of data that lacks in quality. If certain groups of employees see very little representation within a company, and algorithmic models are trained on employees of the company, then the resulting model will not test well on the groups who lack representation (unseen data with respect to the model). The same discriminatory behavior can occur when companies look to promote employees based trained models. Therefore, it is extremely important that the people handling the automated tools used in the hiring process understand the pitfalls of those tools and are completely transparent with regards to their use and the factors used in the them

Risk Assessment Tools

COMPAS is a risk assessment tool developed by Northpointe (now Equivant) used in the U.S. to assess a defendant's risk of recidivism (the likelihood that the defendant will reoffend when released). ProPublica, an independent, non-profit newsroom, raised questions about the scientific validity of COMPAS, stating that the risk assessment tool was racially biased against Black defendants. Equivant disputed the claim of racial bias.

To conduct their analysis, ProPublica used data from 10,000 criminal defendants in Broward County, Florida and compared their actual recidivism rate against what COMPAS predicted in the 2 years after the defendants were scored. While the algorithm correctly predicted the recidivism for Black and White defendants at roughly equal rates (59% for White defendants and 63% for Black defendants), ProPublica was interested in how COMPAS incorrectly predicted recidivism at different rates between Black and White defendants. The takeaways from ProPublica's analysis is that "black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent)" and that "white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent)."

In their counterargument against ProPublica, Equivant highlighted that even within ProPublica's findings, the rate of accuracy for COMPAS scores was the about the same for black and white defendants. Thus, the algorithm cannot be biased as its proportionally accurate at the same rates for both groups. This problem of the algorithm being simultaneously fair and unfair began a debate about algorithmic fairness which I discuss in the next section.

While both sides made their arguments about the fairness of the COMPAS tool, the Wisconsin Supreme Court made the decision on the legality of using the tool, focusing more on the legality on the use of risk-assessment instruments that cannot be transparent due to trade secrets. In the court case Loomis v. Wisconsin, Eric Loomis challenged the state's use of a proprietary/non-disclosable software in his sentencing, stating that it violates his right to due process because he cannot challenge the scientific validity and accuracy of the software and thus, cannot determine if race or gender was taken into account in his sentencing. The Wisconsin Supreme Court ruled against Loomis, stating that while the COMPAS cannot be used to make definite decisions, it can be cautiously used as a supplement for a human decision-maker.

I disagree with the ruling for a number of reasons. More specifically, I disagree with allowing the software to remain proprietary. Firstly, the ruling says that the risk assessment tool should be "cautiously used" but how do we know just how much a judge uses a risk score to make their decision on a defendant. Seeing the risk score assigned a defendant will likely influence a judge's decision. Furthermore, ProPublica did show that the false positive rate in identifying defendants as high risk for recidivism was higher for Black defendants than for White defendants. An important question to ask is how will judges interpret the information on the potential inaccuracies of the COMPAS tool? As I stated before, a frequent misconception is that algorithms operate with objectivity so how likely is it that a judge will make a decision that conflicts with an "impartial" algorithm. While I agree with the ruling that judges need to take caution when using COMPAS, I believe that the tool should be completely transparent.

Discussion on Algorithmic Fairness

The debate between ProPublica and Equivant gathered attention from the press and academics, sparking dialogue about the definitions of "fairness" and "bias" and how/if these words could be mathematically defined. The amount of scholarly material discussing these terms from both a philosophical and mathematical perspective deserves its own in-depth paper but for the purposes of this paper, I will provide a surface level overview of the debate.

The ultimate question asked by scholars in the wake of the ProPublica vs Equivant debate was "Since blacks are re-arrested more often than whites, is it possible to create a formula that is equally predictive for all races without disparities in who suffers the harm of incorrect predictions?" The consensus seems to be that it is not possible. An algorithm created to be equally predictive for all races WILL lead to disparities in the false positive rates when the base rates of re-arrests differ between the groups. Put simply, "A risk score, they found, could either be equally predictive or equally wrong for all races — but not both" and a mathematical proof was constructed to show that both definitions of fairness could not be satisfied at the same time. The argument for the alternate definition of algorithmic fairness—one in which the error rates are equal among the different groups—is that no group should suffer from the inaccuracies of an imperfect model. Empirically, the result of changing the definition of what "fair" is in the context of risk assessment is that now Black defendants were predicted accurately at a significantly higher rate than White defendants.

The discussion on algorithmic fairness is an important one to have and the debate ProPublica vs. Equivant debate was a steppingstone in bringing the discussion to light. In the next section, I discuss the approaches that have been suggested in the effort to curb algorithmic bias.

Mitigating Algorithmic Bias

The Technical Approach

The technical approach for mitigating algorithmic bias often comes down to the data quality and quantity. As explained in the case of the risk assessment tool, achieving both predictive parity and equal error rates is impossible due to baseline differences in re-arrest rates. There is an inherent tradeoff that. Understanding the biases in the data and resulting inaccuracies of a model built of that data can at least help make informed decisions when analyzing the outcomes of such a model.

A toolkit exists created by IBM, named AI Fairness 360, that provides "comprehensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models." The explanations of the fairness metrics are an interesting read. IBM even states that the fairness interventions are complicated and are highly susceptible to changes in the data.

International Human Rights Law as a Framework for Algorithmic Accountability

While it is technically challenging to create a "fair" or "unbiased" algorithm, legal approaches can be taken in the effort of holding algorithms accountable. An article published by

Lorna McGregor, Daragh Murray, and Vivian Ng explores how existing international human rights laws can operate as a framework on the actors who hold different responsibilities in the development and usage of an algorithm and as a framework on the entire lifecycle of an algorithm, from conception to deployment.

My only criticism of this approach to algorithmic accountability is that it would be difficult to enforce such a framework when international human rights law is already ignored by a lot of countries to a significant extent. Private corporations would be inclined to ignore such a framework as well if it hurts their bottom line.

Conclusion

As data continues to grow at a rapid rate, the use of algorithms to leverage this data will expand and permeate public and private life. The benefits that come with the use of algorithms—reduced costs, greater efficiency, insights on trends, prediction, etc.—are too worthwhile to ignore. Nevertheless, it's crucial that responsibility is taken in handling these algorithms, especially when it is used in decision-making in contexts where civil liberties can be violated. Algorithmic bias has become a major topic of concern and awareness of it has grown. It will be interesting to see what the future holds for algorithmic bias and if the technical and legal challenges that it introduces can be overcome.

Citations

- *AI Fairness 360 - Resources*, http://aif360.mybluemix.net/resources#overview.
- Angwin, Julia, and Jeff Larson. "Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say." *ProPublica*, 9 Mar. 2019, https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.
- Barkow, Rachel E. "State v. Loomis." *Harvard Law Review*, 10 Mar. 2017, https://harvardlawreview.org/2017/03/state-v-loomis/.
- Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." *Reuters*, Thomson Reuters, 10 Oct. 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
- Larson, Jeff, et al. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, 9 Mar. 2019, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
- Lum, Kristian, and William Isaac. "Royal Statistical Society Publications." *Royal Statistical Society*, John Wiley & Sons, Ltd (10.1111), 7 Oct. 2016, https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x.
- "Response to ProPublica: Demonstrating Accuracy Equity and Predictive Parity." *Equivant*, 20 Mar. 2019, https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/.