Machine Learning of Criminal Justice in the US

CPSC 490

Student number: 15787154

Mar. 27, 2019

Almost everyone in North America have been subjected to a computerized predictive algorithm of the following form: if you are like X, you may like Y. Movie, music streaming services, as well as online shopping platforms are regularly looking at the pattern of your behaviours, comparing that with past customers' behavioral patterns, and trying to make predictions for you. Financial institutions such as insurance company and banks use computer algorithm to make decisions on whether one gets a loan or a lower insurance payment. Moreover, in recent years, governments have adopted machine algorithms and implemented them in criminal justice systems; courts are routinely using predictive models to aid judges to conduct more accurate and efficient trials. However, studies have shown that such algorithms are sometimes subjected to ethical implications such as racial bias and gender bias. In this paper, I am going to explain how predictive algorithm used in U.S. criminal justice system works and explore the ethical implications with one famous court case of Eric Loomis, where he was given an unfair sentence for his crime, as well as a research study conducted by Professor Hany from Dartmouth College. Finally, I will provide recommendations to mitigate and prevent unfairness in criminal justice system in all of individual, business, and government levels.

One of the most widespread algorithm applied in criminal justice system is at the point of arrest, where judges make bail decisions by referencing the risk score of each convict. Others might think it is a black box with complicated algorithms; however, it is a rather simple idea. A defendant who has been charged with a crime has his or her information extracted and inputted into a computer algorithm, and that algorithm outputs a risk factor that quantifies the likelihood of that person committing a crime in the future. If one is assessed to be high-risk, then the judge may deny bail and he or she will be held in prison awaiting trial;

whereas if one is assessed to be low-risk, then he or she may be released for pending trial. U.S. courts adopted a similar computer algorithm called COMPAS to assess the likelihood of a defendant becoming a recidivist. There are three types of scales that COMPAS uses to make this assessment: Pretrial Release Risk Scale, General Recidivism Scale, and Violent Recidivism Scale. All of them are "designed using behavioral and psychological constructs of very high relevance to recidivism" (Wikipedia, 2019). The Violent Recidivism Scale is calculated based on a range of factors related to the convict, and the score is computed by this equation: "Violent Recidivism Risk Score = (age * −w) + (age-at-first-arrest * −w) + (history of violence * w) + (vocation education * w) + (history of noncompliance * w), where w is weight, the size of which is determined by the strength of the item's relationship to person offense recidivism that we observed in our data" (Wikipedia, 2019), and it is always adjusting in attempt to minimize the error of the model. However, many court cases as well as researches show that COMPAS has racial and gender bias.

Take the example of a famous court case of Eric Loomis, who was pleaded guilty for attempting to flee an officer and operating a vehicle without the owner's consent. Even though neither of his crime mandates the prison time, he was "prescribed a 11-year sentence: six years in prison with five years of extended supervision"(Israni, 2017) because he scored a high risk of recidivism as predicted by COMPAS. No one knows exactly how the other two COMPAS risk scales are calculated, as Northpointe Inc. refuses to disclose the proprietary algorithm. Mr. Loomis challenged the use of COMPAS, but the Wisconsin Supreme Court rejected his challenge. Moreover, Mr. Loomis was not allowed to audit the algorithm even though he suspected that the algorithm has existing bias that might have led to unfair

sentencing. Besides COMPAS displaying a tendency towards gender bias, there are statistics and research done to show that not only does it contain gender bias, but it also has racial bias.

In 2016, investigative journalists from ProPublica published a troubling report on COMPAS. They found this particular algorithm was significantly disadvantageous to black defendants: if a black individual is subjected to this algorithm, he or she is more likely to be assessed as high-risk and incarcerated than if he or she is white. But how is this possible? How are we allowing software to make life altering decisions on people when they have these types of inequalities? Humans build algorithms to make decisions for us with the intention to eliminate man-made error and bias and make assessments more accurate, more objective, and more fair. However, when Professor Farid took initiative to examine the ethical concerns of COMPAS, he found that even though the algorithm doesn't have race as an explanatory variable, the outcome is still racially biased. He performed another online human survey similar to the structure of COMPAS and found that human decisions have the similar accuracy as COMPAS. We need to uncover how humans and the software have a racial bias when no information about the race is given, and how random people on the internet are as good as a commercial software that is being deployed. To do that, we have to answer two questions: what information is in the data which the algorithm is built upon, and how is it operating algorithmically.

In the case of COMPAS, Professor Farid spotted an algorithm bias against black individuals, but as he conducted more experimental trials and tested different explanatory variables, he found the bias originated from the data that was fed to the algorithm. If we take a closer look at NorthPointe's *Practitioner's Guide to COMPAS Core;* the documentation

points out that one important explanatory variable used for calculating risk score is number of prior records made in the past, which it might seem like an unbiased variable as it is positively correlated with the magnitude of risk score. However, in United States, black individuals are more likely to be arrested or charged with a crime, as statistics have shown that black people in New York City are eight times more likely to be arrested for marijuana than white people ; and in California, report have shown that black people are three times more likely to be arrested for crime. Because of the inequity in the criminal justice system, race is slipping into the classifier unknowingly to the general public. By reverse engineering the COMPAS algorithm, Professor Hany has identified the problem that is causing racial bias, and we can conclude that the number of crime is a proxy for race. What can individuals do to protect themselves from being racially discriminated by COMPAS? And how can the government ensure the source of data is unbiased?

Citizens should fight for the right to audit or challenge the technology that are used on them if they believe there is injustice. Similar to GDPR in Europe that grants people the control their personal data, if Canadian and American government want to use algorithms to make life altering decisions on people, they should also give people the right to defend themselves in case of unfair decision making. On the other hand, the government should raise the awareness of implicit bias—in this case is racial bias—among all police force by routinely bringing up the conversation between community and the police, as well as implementing policies and trainings to reduce the impact of bias. On the business level, NorthPointe should follow the three principles when constructing their algorithm. From start to finish, human augmentation should be required and conducted by full-spectrum teams with diverse individuals that can check each other's blind spots. Bias evaluation should be done

not only at the stage of completion but also at the stage of initial data collection. Lastly, building trust within stakeholders is not only done through informing what data is being held, but also setting up policies to protect stakeholders' privacy. By abiding to these principles, tech companies can build responsible machine learning algorithms and improve the accuracy of human decision making.

In the case of COMPAS, the likelihood of recidivism computed using machine learning algorithm does not yield a greater accuracy than humans. The mistake that the developers of COMPAS made is that they failed to treat the bias within the training data, which was carried over to the model, then the risk scores, resulting in unfair sentencing. We as software developers and data scientists must realize that artificial intelligence and machine learning are not inherently less biased than humans. A biased algorithm does not make better decisions than humans; this is particularly true when the data that are being fed into these algorithms mirror existing social inequalities as the biased results might run into a constant feedback loop that amplifies bias, which can greatly backlash our initial intentions for implementing the algorithm. To reduce bias within the algorithm, the tech industry should come together and set up bias eliminating guidelines for developers to follow; and as stakeholders and users, we should be aware of the information that are collected from us, as well as fight for the right to have full control over our personal data.

Bibliography

Corbett-Davies, Pierson, Feller, Sharad. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. Retrieved from https://www.washingtonpost.com

Dressel, Julia & Farid, Hany. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances. 4. eaao5580. 10.1126/sciadv.aao5580.

Keesee, Tracie. L. (2015). Three Ways to Reduce Implicit Bias in Policing. Retrieved from https://greatergood.berkeley.edu/

NorthPointe. (2015). Practitioner's Guide to COMPAS Core. https://assets.documentcloud.org/

Wikipedia. (2019). COMPAS (software). Retrieved from https://en.wikipedia.org/