Algorithmic Fairness in Recidivism

Dhruval Bhatt
April 19th, 2020

Constructing and choosing the best model requires robust inputs and a well-defined criterion for evaluate the model's performance. In most predictive models it is common to select the model with highest accuracy. However, when algorithms are used to make or influence decisions that impact human lives, a careful assessment of sources of bias and fairness of algorithmic output is needed. In the criminal justice system, algorithms are increasingly used by judges to help evaluate a convict's pretrial, parole, and sentencing conditions, that certainly has long term impacts on the person if the decision is incorrect or biased. Here, we consider how to construct the best model that predicts the likelihood of a convicted felon committing a new felony at some point in the next three years.

At the onset, it is important to define the problem and set an evaluation criterion for accuracy and fairness. The accuracy of a classification model can be determined by proportion of misclassification. These two errors possible are false positives (classified as high risk when they are not) and false negatives (classified as low risk when they are a risk). This is a punitive model (consequence is a punishment) so "one might make a case for caring more about False Positives to avoid punishing the innocent based on the output of the model" (Pandey 2019). On the other hand, if a likely repeat offender is categorized as low risk and allowed to be free in society without the appropriate retribution, there could be additional crimes that could affect law abiding citizens. Therefore, the core goal should be to minimize both scenarios overall but also ensure that there is no algorithmic bias that systematically misclassifies a subgroup. That is, if one race is usually misclassified as high risk and the opposite for another - a problem that was flagged in COMPAS's recidivism algorithm.

In considering the inputs, the age-old warning in computing, "garbage in, garbage out" (GIGO), applies for such a case more than ever. Many learning models are trained on inputs from historical data, so it is important to evaluate if the input is not inherently biased. Criminal justice system data represents people that were arrested and charged for a crime but not all crime. Humans in criminal justice system are susceptible to explicit and implicit biases that might lead to racially profiled (or gendered) arrests leading to a dataset that has higher representation of one race (or

gender) over another. It is important to understand the distribution of data and correct for or account for imbalances in data before the training the model. This is important for immediate output as well as long term institutional bias as, “using biased data or algorithms for further data collection, [forms] a pernicious feedback loop that can amplify discrimination over time” (Kearns et. al, 2020, 114)

Additionally, it is important to select appropriate attributes as input. Kearns and Roth point out that simply removing the racial or gender information seems like an easy solution for fairness in input, but it reduces the accuracy and since race/gender is usually correlated with other information, the algorithm is still likely to be biased. If several attributes are available for training a model, methods such as LASSO can be to reduce inputs to those that have an impact on the model and improve model accuracy. In addition, theoretical work in sociology and criminal psychology should be consulted to add or remove attributes. Even in a data driven approach, it is important to remember that correlation is not causation. For instance, most criminals may be poor but most poor people are not criminals. Just relying on data could unknowingly propagate fallacies embedded in the data available. Northpointe’s approach to include philosophical questions as a part of the “137 questions in their core product” seems like a good start but they should be refined to reflect the true understanding of an individual’s remorse and impulses over general stereotypes (Angwin et.al, 2016).

Furthermore, to ensure algorithmic fairness methods such as statistical parity, balancing false positives and balancing false negatives. In this situation, it is best to balance the classes as they “simply ask that once we condition on the “correct” answer for a person, the chance of making a mistake on them should not depend on which group they belong to” (Kleinberg et. al, 2016, 3). In any case, with such a constraint, a human need to decide on tradeoffs in accuracy or fairness method as not all of it can be optimized. The recidivism algorithm should be tuned to incorporate fairness at the cost of accuracy, but the judge should use their own experience and real-time input to evaluate the accuracy of algorithmic output. It would be helpful that along with a risk assessment score, the algorithm could provide quantitative confidence intervals for the expected accuracy of the decision. Such a two-tiered decision-making model would be ideal in this case as the algorithm helps guide the decision based on common factors and past predictors but gives room for humans

to provide an intuitive decision to determine if a person is earnest in their commitment to a better life or not.

While it is important to understand the limitations of machine learning in social context, the merits of algorithmic output should not be ignored. Just like humans, algorithms can be biased but "with the appropriate requirements in place, algorithms create the potential for new forms of transparency and hence opportunities to detect discrimination that are otherwise unavailable" (Kleinberg et al, 2019, 113). The biases revealed in algorithms could in fact be a mirror to our institutional discrimination and implicit bias. In some cases, monitoring and regulating human behavior is difficult but with the right judgements and tuning, an algorithm can be trained to produce transparent and equitable results.

Words: 963

References

Angwin, Julia, et al. "Machine Bias." ProPublica, 23 May 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Kearns, M., & Roth, A. (2020). The ethical algorithm: the science of socially aware algorithm design. New York, NY: Oxford University Press.

Kleinberg, Jon, et al. "Discrimination in the Age of Algorithms." 2019, doi:10.3386/w25548.

Kleinberg, Jon, et al. "Inherent Trade-Offs in the Fair Determination of Risk Scores" 2016, https://arxiv.org/abs/1609.05807v2

Pandey, Parul. "Is Your Machine Learning Model Biased?" Medium, Towards Data Science, 6 Aug. 2019, towardsdatascience.com/is-your-machine-learning-model-biased-94f9ee176b67.