Opinion | Trusting Algorithms with our Lives: How do we design fair algorithms in a biased criminal justice system?

By Emilee Fulton and Javin White
Edited by Carlos Garcia and Layla Oesper
Fulton, White and Garcia worked alongside fellow undergraduate researchers Kellen Dorschen, Cameron Kline-Sharpe, and Dillon Lanier to produce the results found in this study. All research was conducted under the guidance of Dr. Layla Oesper, Assistant Professor of Computer Science at Carleton College.

March 11, 2020

Today, all 50 states use pre or post trial risk assessment tools to help judges determine sentence length or probation time of individuals with criminal records. These tools use a variety of information such as history of violent crime and failure to appear in court, to predict the likelihood that a convicted individual will re-offend in the next two years. While the exact algorithms used to predict recidivism varies by U.S. county, a question that both law enforcement and criminal justice rights activists grapple with is: how and to what extent should computer algorithms be used when determining sentences and probation?

Courtrooms demand unbiased sentencing for those being put on trial, however, judges are not always akin to making unbiased decisions. It is for this reason that statistical prediction tools serve as an asset to our current judicial system. Unfortunately, even when computer algorithms are based solely on data, it does not mean that they are inherently fair.

Currently, there is no nationwide standard for risk-assessment algorithms. Fifteen states use internally developed algorithms, while others use commercial algorithms. One such commercial algorithm, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), uses 137 features, including past criminal record and responses to a personality test, to predict the likelihood that an individual will reoffend within two years.

Northpointe, the company that developed COMPAS, conducted their own "in-depth analysis" of the COMPAS risk scales to prove the efficacy of the algorithm. However, a 2016 study by ProPublica revealed that the algorithm displays racial bias in its assessment of future criminal behavior. Specifically, Probublica obtained criminal record data containing 17 features from Broward County, Florida via a public records request. From this dataset they found that while COMPAS has an accuracy of 65% (where 50% accuracy is equal to random guessing), the algorithm tends to falsely label 44.9% of African-Americans as being at a higher risk to reoffend while only falsely labeling 23.5% of White Americans as being at a higher risk to reoffend. Conversely, White Americans are falsely predicted to have a low risk of reoffending at about twice the rate of African-Americans.

Mediating these false predictions is not easy, but it is necessary to do so given the usage implications of these tools in court proceedings. Given that African-Americans make up over 50% of the US prison population but only 14% of the total US population, our group of six researchers aimed to develop algorithms that are more accurate and more fair than COMPAS. We define "more accurate" as having a higher rate of true positives and true negatives than COMPAS. Similarly, we define "more fair" as minimizing the rate of false positives, the number of people who were predicted to reoffend but did not.

Before designing any algorithms, we began by researching previous studies on risk assessment. We found that while the inaccuracies of risk assessment algorithms, such as COMPAS, are known, it is difficult for researchers to improve existing tools because the criminal record data required to test new algorithms is difficult to obtain. Police report data can be requested individually by U.S. county, however these requests are usually for individual incidents, and the requestor must provide specific information about each incident.

Due to the sparsity of publically available data, much of the research conducted in the field of risk assessment algorithms, including ProPublica's 2016 expose on COMPAS, uses the same dataset from Broward County Florida, containing information on 11,000 convicted individuals. While criminal conviction data is undoubtedly sensitive, if all work in the field is conducted using the same 11,000 data points, we run the risk of research towards improving risk assessment algorithms becoming generalized to this specific dataset. With this problem in mind, we searched for another option.

We soon found that we could request the Post Conviction Risk Assessment (PCRA) dataset from the U.S. Courts, which contains about 300,000 individual federal cases. However upon requesting the data, we were met with two months of silence. Finally we received an email denying our request for data, stating our project goal of defining "fair" risk assessment algorithms is "not currently a research priority at the Administrative Office of the U.S. courts."

Given the limited number of options for acquiring data, we had to make due with the data from Broward County, Florida. We used this dataset to train and test three different classification algorithms.

First, we developed a Naive Bayes algorithm to predict the likelihood that an individual will reoffend. Naive Bayes is a binary classification algorithm that uses Bayes' theorem and the naive assumption that all features are independent, to determine probability of an event occuring. Note that in the context of this problem, features are pieces of information about someone, such as their age, sex, and previous charges. Our Naive Bayes algorithm accurately predicted whether an individual would reoffend 66% of the time, a 1% improvement over COMPAS, which has an accuracy of 65% on the Broward County dataset.

Upon further investigation of the increase in accuracy over COMPAS, we found a problem. Specifically, because in the Broward County dataset it is less likely that someone from any

demographic will commit another crime, than that they will not, our Naive Bayes algorithm was just a fancy way of predicting that nobody will ever reoffend; our 66% "accuracy" was the rate of not reoffending in the dataset.

In hopes of developing a smarter classification algorithm, we turned to decision trees. Decision trees are classifiers that group items into different classes by repeatedly splitting the items based on features of the data. Our tree split the data on features such as the number of juvenile crimes people had committed and age. Our Decision Tree algorithm boasted an accuracy of 68%, a 3% improvement over COMPAS.

Finally, using the Keras library, we implemented an artificial neural network algorithm. A neural network is composed of hidden layers that use the input features to output a decision. We applied the decision the network produced to a prediction about each individual in the dataset. Our neural network algorithm was accurate around 70% of time.

So which algorithm is best? While there is no single "right" answer, we believe our decision tree algorithm is best in the context of predicting recidivism. While all three algorithms were more accurate than COMPAS, Naive Bayes never predicted that anyone would recidivate, making it not useful as a risk assessment tool. Additionally, while the Neural Networks algorithm boasted the best accuracy, given the obscure nature of neural network algorithms the question of how the algorithm actually used the given features about each individual to predict recidivism was not clear. Conversely, the Decision Tree algorithm clearly showed how each specific feature was weighed in predicting whether an individual will reoffend. As the goal of our research was to make risk assessment tools more transparent and fair, the decision tree algorithm best suited these needs.

While endless research can be done to determine the "most fair" risk assessment algorithm, if a "fair" algorithm is misused in practice, then it once again becomes a problem. With this in mind, in 2019, over two hundred justice and human rights organizations signed, "A Shared Statement of Civil Rights Concerns," a report that outlines six guiding principles for use of pretrial risk assessment tools in the United States. Overall the statement stresses that risk assessment algorithms should only be used as *one* piece of a carefully made decision. Additionally, the statement proclaims that risk assessment algorithms can provide evidence for the exoneration of an individual, but not for their detention.

By following this guideline, the United States judicial system could recommend the release of low risk individuals, thus speeding up the costly and lengthy trial process, while posing little risk to society at large. Further, those categorized as high risk to reoffend would go through the traditional trial process, which would not be swayed by the results of the risk assessment algorithm.

Additionally, the input data to risk assessment algorithms should be considered. In an ideal world, risk assessment algorithms would not use racially charged features like education level,

number of suspensions and expulsions received, and financial stability. Instead, risk assessment algorithms should be more like our decision tree: open-source, transparent in their calculations, and only considering features that directly relate to the judicial system, such as their number of prior convictions and failures to appear in court. If this could be achieved, risk assessment tools have the potential to decrease time between arrest and conclusion of a case, easing tensions on bogged down court systems across the country.