Opening the lid on criminal sentencing software

July 20 2017, by Robin A. Smith

The government shouldn't have to pay private companies to predict recidivism, using mysterious computer formulas that are tightly guarded as trade secrets -- at least not without proof that such models work best, says Duke professor Cynthia Rudin. Credit: Duke University

In 2013, a Wisconsin man named Eric Loomis was convicted of fleeing

an officer and driving a car without the owner's consent.

He was denied probation and sentenced to six years in prison based, in part, on a prediction made by a secret computer algorithm.

The algorithm, developed by a private company called Northpointe, had determined Loomis was at "high risk" of running afoul of the law again. Car insurers base their premiums on the same sorts of models, using a person's driving record, gender, age and other factors to calculate their risk of having an accident in the future.

Loomis challenged the sentencing decision, arguing that the algorithm's proprietary nature made it difficult or impossible to know why it spat out the result it did, or dispute its accuracy, thus violating his rights to due process.

The state supreme court ruled against him in 2016, and this June the U.S. Supreme Court declined to weigh in.

But there are good reasons to remain wary, says Cynthia Rudin, associate professor of computer science and electrical and computer engineering at Duke University.

Every year, courts across the country make decisions about who to lock up, and for how long, based on "black box" software whose opaque inner workings are a mystery—often without proof that they're as good or better than other tools.

Defenders of such software say black box models are more accurate than simpler "white-box" models that people can understand.

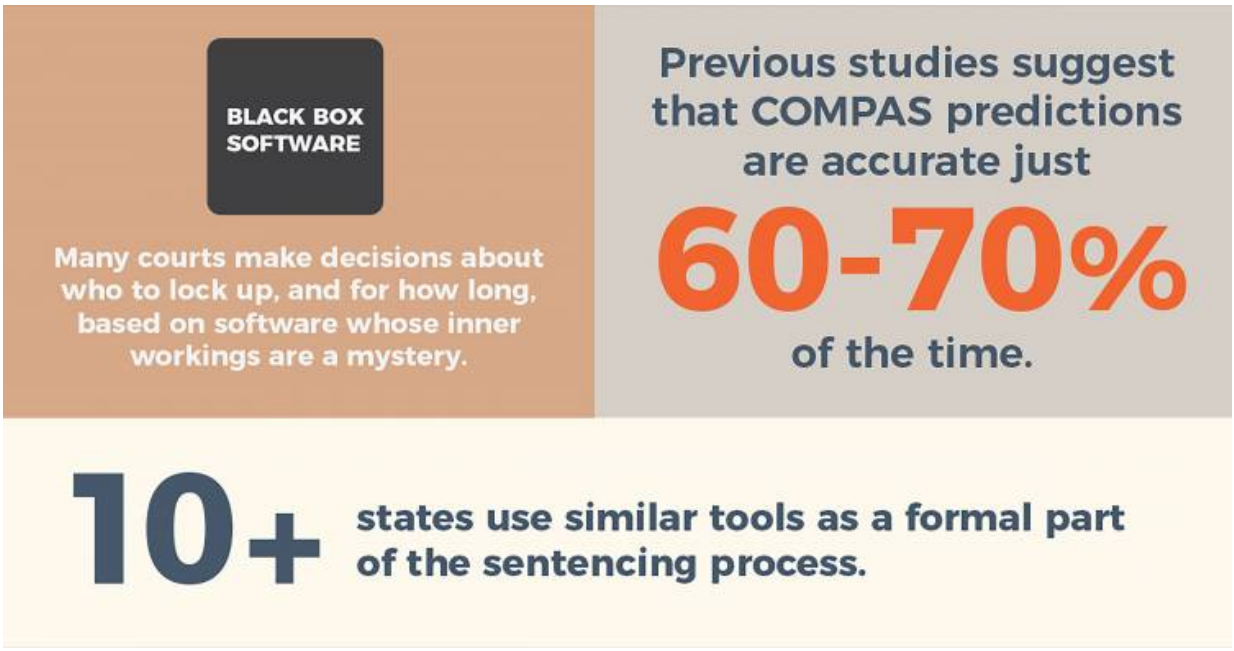
But Rudin says it doesn't have to be this way.

Using a branch of computer science called machine learning, Rudin and colleagues are training computers to build statistical models to predict future criminal behavior, called recidivism, that are just as accurate as black-box models, but more transparent and easier to interpret.

Recidivism forecasts are not new. Since the 1920s, the U.S. criminal justice system has used factors such as age, race, criminal history, employment, school grades and neighborhood to predict which former inmates were most likely to return to crime, and to determine their need for social services such as mental health or substance abuse treatment upon release.

Northpointe's tool, called COMPAS, relies on a person's criminal record, age, gender, and answers to dozens of questions about their marital and family relationships, living situation, school and work performance, substance abuse and other risk factors. It then uses that information to calculate an overall score that classifies an offender as low, medium or high risk of recidivism.

Similar tools are a formal part of the sentencing process in at least 10 states.



Credit: Duke University

Proponents say the tools help the courts rely less on subjective intuition and make evidence-based decisions about who can safely be released instead of serving jail time, thus reducing prison overcrowding and cutting costs.

But just because a risk score is generated by a computer doesn't make it fair and trustworthy, Rudin counters.

Previous studies suggest that COMPAS predictions are accurate just 60 to 70 percent of the time. In independent tests run by ProPublica, researchers analyzed the scores and found that African Americans who did not commit further crimes were nearly two times more likely than whites to be wrongly flagged as "high risk." Conversely, whites who became repeat offenders were disproportionately likely to be misclassified as "low risk."

COMPAS isn't the only recidivism prediction tool whose validity has been called into question.

With any black box model, it is difficult to tell whether the predictions are valid for an individual case, Rudin says. Errors could arise from inaccurate or missing data in a person's profile, or problems with the data the models were trained on. Models developed based on patterns in data from one state or jurisdiction may not do as well in another.

Under the current system, even simple data entry errors can mean inmates are denied parole. The algorithm just crunches the numbers it's given; there's no recourse.

"People are getting different prison sentences because some completely opaque algorithm is predicting that they will be a criminal in the future," Rudin says. "You're in prison and you don't know why and you can't argue."

Rudin and her colleagues are using machine learning to make it possible for offenders to ask why.

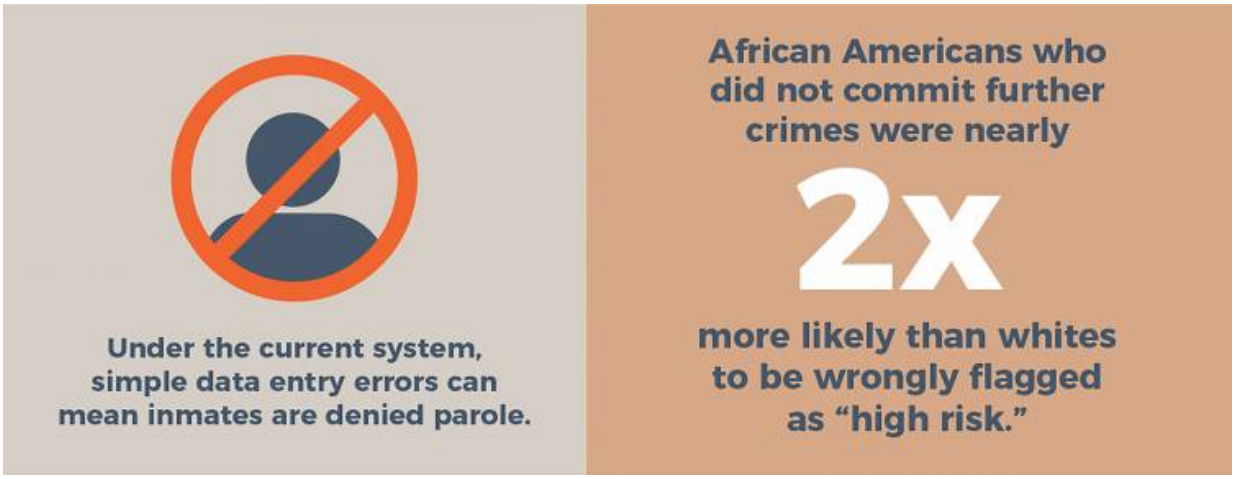
In one recent study, Rudin and collaborators Jiaming Zeng, a graduate student at Stanford University, and Berk Ustun, a graduate student at MIT, describe a method they developed, called Supersparse Linear Integer Model, or SLIM.

Using a public dataset of over 33,000 inmates who were released from prison in 15 states in 1994 and tracked for three years, the researchers had the algorithm scan the data to look for patterns. The system took into account things like gender, age, criminal history and dozens of other variables, trying to find ways predict future offenses. It then built a model to predict whether a defendant will relapse or not, based on those same rules.

"For most machine learning models, the formula is so big it would take more than a page to write it down," Rudin said.

Not so with the SLIM method. Judges could use a simple score sheet small enough to fit on an index card to turn the results of the SLIM model into a prediction.

All they have to do is add up the points for each risk factor and use the total to assign someone to a category. Being 18 to 24 years old adds two points to a person's score, for example, as does having more than four prior arrests.



Credit: Duke University

The SLIM method lets users make quick predictions by hand, without a calculator, and gauge the influence of different input variables on the result.

The algorithm also builds models that are highly customizable. The

researchers were able to build separate models to predict the likelihood of arrest for different crimes such as drug possession, domestic violence or manslaughter. SLIM predicted the likelihood of arrest for each crime just as accurately as other machine learning methods.

The SLIM method could also be applied to data from different geographic areas to create customized models for each jurisdiction, instead of the "one size fits all" approach used by many current models, Rudin says.

As for transparency, the models are built from publicly available data sets using open-source software. The researchers disclose the details of their algorithms, rather than keeping them proprietary. Anyone can inspect the data fed into them, or use the underlying code, for free.

In a new study, Rudin and colleagues introduce another machine learning algorithm, called CORELS, that takes in data about new offenders, compares them to past offenders with similar characteristics, and divides them into "buckets" to help predict how they might behave in the future. Developed with Elaine Angelino, a postdoctoral fellow at the University of California, Berkeley, Harvard computer science professor Margo Seltzer and students Nicholas Larus-Stone and Daniel Alabi, the model stratifies offenders into risk groups based on a series of "if/then" statements.

The model might say, for example, that if a defendant is 23 to 25 years old and has two or three prior arrests, they are assigned to the highest risk category, 18- to 20-year-olds are in the second highest risk category, men aged 21-22 are next, and so on.

The researchers ran their algorithm on a dataset of more than 11,000 defendants in Florida, and compared the recidivism rates predicted by the CORELS algorithm with the arrests that actually occurred over two

years. When it comes to differentiating between high- and low-risk offenders, the CORELS approach fared just as well or better than other models, including COMPAS.

But unlike those models, CORELS makes it possible for judges and defendants to scrutinize why the algorithm classifies a particular person as high or low risk, Rudin says.

None of the research team's models rely on race or socioeconomic status.

The researchers will present their approaches at the 23rd SIGKDD Conference on Knowledge Discovery and Data Mining, held in Halifax, Nova Scotia, Aug. 15-17.

The stakes in the criminal justice system are too high to blindly trust in black box algorithms that haven't been properly tested against available alternatives, Rudin says.

Next year, the European Union will begin requiring companies that deploy decision-making algorithms that significantly affect EU citizens to explain how their algorithms arrived at their decisions.

Rudin says the technical solutions are already out there that would enable the criminal justice system in the United States or anywhere else to do the same, at considerably less cost—we just have to use them.

"We've got good risk predictors that are not black boxes," Rudin says. "Why ignore them?"

More information: Jiaming Zeng et al. Interpretable classification models for recidivism prediction, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (2016). DOI: 10.1111/rssa.12227

Learning Certifiably Optimal Rule Lists for Categorical Data. *arXiv*. arxiv.org/abs/1704.01701