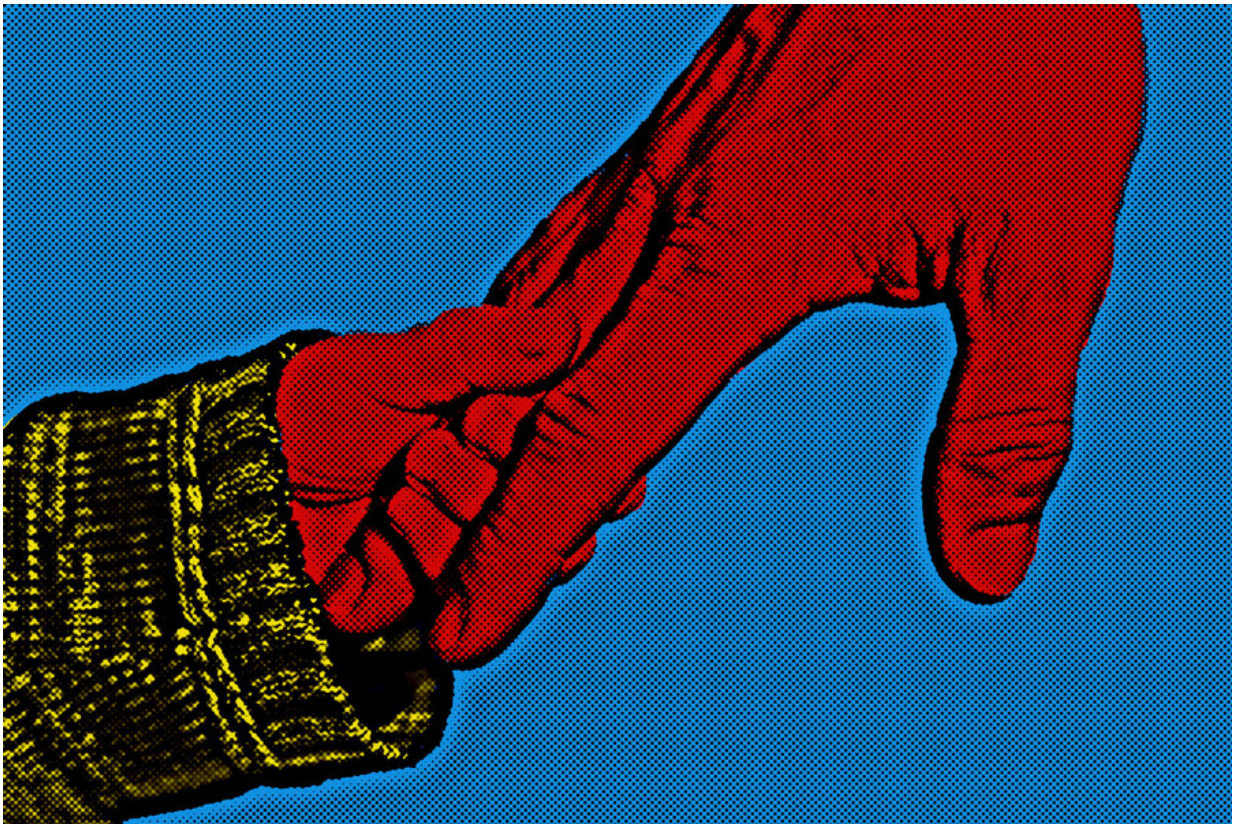
Making machine learning more useful to high-stakes decision makers

October 28 2021, by Adam Zewe



A new visual analytics tool helps child welfare specialists understand machine learning predictions that can help them make decisions. Credit: Christine Daniloff, MIT

The U.S. Centers for Disease Control and Prevention estimates that one

in seven children in the United States experienced abuse or neglect in the past year. Child protective services agencies around the nation receive a high number of reports each year (about 4.4 million in 2019) of alleged neglect or abuse. With so many cases, some agencies are implementing machine learning models to help child welfare specialists screen cases and determine which to recommend for further investigation.

But these models don't do any good if the humans they are intended to help don't understand or trust their outputs.

Researchers at MIT and elsewhere launched a research project to identify and tackle machine learning usability challenges in child welfare screening. In collaboration with a child welfare department in Colorado, the researchers studied how call screeners assess cases, with and without the help of machine learning predictions. Based on feedback from the call screeners, they designed a visual analytics tool that uses bar graphs to show how specific factors of a case contribute to the predicted risk that a child will be removed from their home within two years.

The researchers found that screeners are more interested in seeing how each factor, like the child's age, influences a prediction, rather than understanding the computational basis of how the model works. Their results also show that even a simple model can cause confusion if its features are not described with straightforward language.

These findings could be applied to other high-risk fields where humans use machine learning models to help them make decisions, but lack data science experience, says senior author Kalyan Veeramachaneni, principal research scientist in the Laboratory for Information and Decision Systems (LIDS) and senior author of the paper.

"Researchers who study explainable AI, they often try to dig deeper into the model itself to explain what the model did. But a big takeaway from

this project is that these domain experts don't necessarily want to learn what machine learning actually does. They are more interested in understanding why the model is making a different prediction than what their intuition is saying, or what factors it is using to make this prediction. They want information that helps them reconcile their agreements or disagreements with the model, or confirms their intuition," he says.

Co-authors include electrical engineering and computer science Ph.D. student Alexandra Zytek, who is the lead author; postdoc Dongyu Liu; and Rhema Vaithianathan, professor of economics and director of the Center for Social Data Analytics at the Auckland University of Technology and professor of social data analytics at the University of Queensland. The research will be presented later this month at the IEEE Visualization Conference.

Real-world research

The researchers began the study more than two years ago by identifying seven factors that make a machine learning model less usable, including lack of trust in where predictions come from and disagreements between user opinions and the model's output.

With these factors in mind, Zytek and Liu flew to Colorado in the winter of 2019 to learn firsthand from call screeners in a child welfare department. This department is implementing a machine learning system developed by Vaithianathan that generates a risk score for each report, predicting the likelihood the child will be removed from their home. That risk score is based on more than 100 demographic and historic factors, such as the parents' ages and past court involvements.

"As you can imagine, just getting a number between one and 20 and being told to integrate this into your workflow can be a bit challenging,"

Zytek says.

They observed how teams of screeners process cases in about 10 minutes and spend most of that time discussing the risk factors associated with the case. That inspired the researchers to develop a case-specific details interface, which shows how each factor influenced the overall risk score using color-coded, horizontal bar graphs that indicate the magnitude of the contribution in a positive or negative direction.

Based on observations and detailed interviews, the researchers built four additional interfaces that provide explanations of the model, including one that compares a current case to past cases with similar risk scores. Then they ran a series of user studies.

The studies revealed that more than 90 percent of the screeners found the case-specific details interface to be useful, and it generally increased their trust in the model's predictions. On the other hand, the screeners did not like the case comparison interface. While the researchers thought this interface would increase trust in the model, screeners were concerned it could lead to decisions based on past cases rather than the current report.

"The most interesting result to me was that, the features we showed them—the information that the model uses—had to be really interpretable to start. The model uses more than 100 different features in order to make its prediction, and a lot of those were a bit confusing," Zytek says.

Keeping the screeners in the loop throughout the iterative process helped the researchers make decisions about what elements to include in the machine learning explanation tool, called Sibyl.

As they refined the Sibyl interfaces, the researchers were careful to

consider how providing explanations could contribute to some cognitive biases, and even undermine screeners' trust in the model.

For instance, since explanations are based on averages in a database of child abuse and neglect cases, having three past abuse referrals may actually decrease the risk score of a child, since averages in this database may be far higher. A screener may see that explanation and decide not to trust the model, even though it is working correctly, Zytek explains. And because humans tend to put more emphasis on recent information, the order in which the factors are listed could also influence decisions.

Improving interpretability

Based on feedback from call screeners, the researchers are working to tweak the explanation model so the features that it uses are easier to explain.

Moving forward, they plan to enhance the interfaces they've created based on additional feedback and then run a quantitative user study to track the effects on decision making with real cases. Once those evaluations are complete, they can prepare to deploy Sibyl, Zytek says.

"It was especially valuable to be able to work so actively with these screeners. We got to really understand the problems they faced. While we saw some reservations on their part, what we saw more of was excitement about how useful these explanations were in certain cases. That was really rewarding," she says.

More information: Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, Kalyan Veeramachaneni, Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making. arXiv:2103.02071v2 [cs.HC], arxiv.org/abs/2103.02071