Harvard Data Science Review • 1.2

Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From An Explainable AI Competition

Cynthia Rudin, Joanna Radin

Published on: Nov 22, 2019

Updated on: Dec 03, 2019

DOI: 10.1162/99608f92.5a8a3a3d

ABSTRACT

In 2018, a landmark challenge in artificial intelligence (AI) took place, namely, the Explainable Machine Learning Challenge. The goal of the competition was to create a complicated black box model for the dataset and explain how it worked. One team did not follow the rules. Instead of sending in a black box, they created a model that was fully interpretable. This leads to the question of whether the real world of machine learning is similar to the Explainable Machine Learning Challenge, where black box models are used even when they are not needed. We discuss this team's thought processes during the competition and their implications, which reach far beyond the competition itself.

Keywords: interpretability, explainability, machine learning, finance

In December 2018, hundreds of top computer scientists, financial engineers, and executives crammed themselves into a room within the Montreal Convention Center at the annual Neural Information Processing Systems (NeurIPS) conference to hear the results of the Explainable Machine Learning Challenge, a prestigious competition organized collaboratively between Google, the Fair Isaac Corporation (FICO), and academics at Berkeley, Oxford, Imperial, UC Irvine, and MIT. This was the first data science competition that reflected a need to make sense of outcomes calculated by the black box models that dominate machine learning–based decision making.

Over the last few years, the advances in deep learning for computer vision have led to a widespread belief that the most accurate models for any given data science problem must be inherently uninterpretable and complicated. This belief stems from the historical use of machine learning in society: its modern techniques were born and bred for low-stakes decisions such as online advertising and web search where individual decisions do not deeply affect human lives.

In machine learning, these black box models are created directly from data by an algorithm, meaning that humans, even those who design them, cannot understand how variables are being combined to make predictions. Even if one has a list of the input variables, black box predictive models can be such complicated functions of the variables that no human can understand how the variables are jointly related to each other to reach a final prediction.

Interpretable models, which provide a technically equivalent, but possibly more ethical alternative to black box models, are different—they are constrained to provide a better understanding of how predictions are made. In some cases, it can be made very clear how variables are jointly related to form the final prediction, where perhaps only a few variables are combined in a short logical statement, or using a linear model, where variables are weighted and added together. Sometimes interpretable models are comprised of simpler models put together (decomposable), or other constraints are put on the model to add a new level of insight. Most machine learning models, however, are not designed with interpretability constraints; they are just designed to be accurate predictors on a static dataset that may or may not represent how the model would be used in practice.

The belief that accuracy must be sacrificed for interpretability is inaccurate. It has allowed companies to market and sell proprietary or complicated black box models for high-stakes decisions when very simple interpretable models exist for the same tasks. As such, it allows the model creators to profit without considering harmful consequences to the affected individuals. Few question these models because their designers claim the models need to be complicated in order to be accurate. The 2018 Explainable Machine Learning Challenge serves as a case study for considering the tradeoffs of favoring black box models over interpretable ones.

Prior to the winners of the challenge being announced, the audience—consisting of power players in the realms of finance, robotics, and machine learning—were asked to engage in a thought experiment where they had cancer and needed surgery to remove a tumor. Two images were displayed on the screen. One image depicted a human surgeon, who could explain anything about the surgery, but had a 15% chance of causing death during the surgery. The other image showed a robotic arm that could perform the surgery with only a 2% chance of failure. The robot was meant to simulate a black box approach to artificial intelligence (AI). In this scenario, total trust in the robot was required; no questions could be asked of the robot, and no specific understanding of how it came to its decisions would be provided. The audience was then asked to raise a hand to vote for which of the two they would prefer to perform life-saving surgery. All but one hand voted for the robot.

While it may appear obvious that a 2% chance of mortality is better than a 15% chance of mortality, framing the stakes of AI systems in this way obscures a more fundamental and interesting consideration: *Why must the robot be a black box?* Would the robot lose its ability to perform accurate surgery if it was enabled with an ability to explain itself? Wouldn't having better communication between the robot and the patient, or a physician, improve patient care rather than diminish it? Wouldn't the patient need to be able to explain to the robot that they had a blood clotting disorder before the surgery?

This possibility, that the robot did not need to be a black box, was not presented as an option, and the audience of the workshop was given only the choice between the accurate black box and the inaccurate glass box. The audience was not told how accuracy was being measured for the surgical outcomes (on what population was the 2% and 15% measured?) nor were they told about potential flaws in the dataset that was used to train the robot. In assuming that accuracy must come at the cost of interpretability (the ability to understand why the surgeon does what they do), this mental experiment failed to consider that interpretability might not hurt accuracy. Interpretability might even improve accuracy, as it permits an understanding of when the model, in this case a robotic surgeon, might be incorrect.

Being asked to choose an accurate machine or an understandable human is a false dichotomy. Understanding it as such helps us to diagnose the problems that have resulted from the use of black box models for high-stakes decisions throughout society. These problems exist in finance, but also in healthcare, criminal justice, and beyond.

Let us give some evidence that this assumption (that we must always sacrifice some interpretability to get the most accurate model) is wrong. In the criminal justice system, it has been repeatedly demonstrated (Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2018; Tollenaar & van der Heijden, 2013; Zeng, Ustun, & Rudin, 2016) that complicated black box models for predicting future arrest are not any more accurate than very simple predictive models based on age and criminal history. For instance, an interpretable machine learning model for predicting rearrest created in work by Angelino et al. (2018), considers only a few rules about someone's age and criminal history. The full machine learning model is as follows: if the person has either >3 prior crimes, or is 18–20 years old and male, or is 21–23 years old and has two or three prior crimes, they are predicted to be rearrested within two years from their evaluation, and otherwise not. While we are not necessarily advocating to use this particular model in criminal justice settings, this set of rules is as accurate as the widely used (and proprietary) black box model called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), as used in Broward County, Florida (Angelino et al., 2018).

The simple model above is also as accurate as many other state-of-the-art machine learning methods (Angelino et al., 2018). Similar results were found across machine learning methods applied to many different types of rearrest prediction problems on other datasets: the interpretable models (which were very small linear models or logical models in these studies) performed just as well as the more complicated (black box) machine learning models (Zeng et al., 2016). There does not seem to be evidence of a benefit from using black box models for criminal risk prediction. In fact, there may be disadvantages in that these black boxes are more difficult to troubleshoot, trust, and use.

There also does not seem to be a benefit in accuracy for black box models in several healthcare domains and across many other high-stakes machine learning applications where life-altering decisions are being made (e.g., Caruana et al., 2015; Razavian et al., 2015; Rudin & Ustun, 2018, who all show models with interpretability constraints that perform just as well as unconstrained models). On the contrary, black box models can mask a myriad of possible serious mistakes (e.g., see Rudin, 2019). Even in computer vision, where deep neural networks (the most difficult kind of black box model to explain) are the state-of-the-art, we and other scientists (e.g., Chen et al., 2019; Y. Li et al., 2017; L. Li, Liu, Chen, & Rudin, 2018; Ming, Xu, Qu, & Ren, 2019) have found ways to add interpretability constraints to deep learning models, leading to more transparent computations. These interpretability constraints have not come at the expense of accuracy, even for deep neural networks for computer vision.

Trusting a black box model means that you trust not only the model's equations, but also the entire database that it was built from. For instance, in the scenario of the robot and the surgeon, without knowing how the 2% and 15% were estimated, we should question the relevance of these numbers for any particular subpopulation of medical patients. Every reasonably complex dataset we have seen contains imperfections. These can range from huge amounts of missing data (that are not missing at random), or unmeasured confounding, to systematic errors in the dataset (e.g., incorrect coding of drug treatments), to data collection issues that cause the distribution of data to be different than what we originally thought.

One such common issue with black box models in medical settings is data leakage, where some information about the label y sneaks into the variables x in a way that you might not suspect by looking at the titles and descriptions of the variables: sometimes you think you are predicting something in the future but you are only detecting something that happened in the past. In predicting medical outcomes, the machine might pick up on information within doctors' notes that reveal the patients' outcome before it is officially recorded and hence erroneously claim these as successful predictions.

In attempting to reckon with widespread concern about the opacity of black box models, some scientists have tried to offer explanations of them, hypotheses about why they reach the decisions they do. Such explanations usually try to either mimic the black box's predictions using an entirely different model (perhaps with different important variables, masking what the black box might actually be doing), or they provide another statistic that yields incomplete information about the calculation of the black box. Such explanations are shallow, or even hollow, since they extend the authority of the black box rather than recognizing it is not necessary. And sometimes, these explanations are wrong.

For instance, when ProPublica journalists tried to explain what was in the proprietary COMPAS model for recidivism prediction (Angwin et al., 2016), they seem to have mistakenly assumed that if one could create a linear model that approximated COMPAS and depended on race, age, and criminal history, that COMPAS itself must depend on race. However, when one approximates COMPAS using a nonlinear model, the explicit dependence on race vanishes (Rudin, Wang, & Coker, 2019), leaving dependence on race only through age and criminal history. This is an example of how an incorrect explanation of a black box can spiral out of control. Perhaps if the justice system had used only interpretable models (which we and others have demonstrated to be equally as accurate), ProPublica's journalists would have been able to write a different story. Perhaps, for instance, they might write about how typographical errors in these scores occur frequently, with no obvious way to troubleshoot them, leading to inconsistent life-altering decision making in the justice system (see, e.g., Rudin et al., 2019).

But back at the 2018 NeurIPS conference, in the room full of experts who had just chosen the robot over the surgeon, the announcer proceeded to describe the competition. The FICO had provided a home equity line of credit (HELOC) dataset, which contains data from thousands of anonymous individuals, including aspects of their credit history and whether or not the individual defaulted on the loan. The goal of the competition was to create a black box model for predicting loan default, and then explain the black box.

One would assume that for a competition that required contestants to create a black box and explain it, the problem would actually need a black box. But it did not. Back in July of 2018 when the Duke team received the data, after playing with it for only a week or so, we realized that we could effectively analyze the FICO data without a black box. No matter whether we used a deep neural network or classical statistical techniques for linear models, we found that there was less than a 1% difference in accuracy between the methods, which is within the margin of error caused by random sampling of the data. Even when we used machine learning techniques that provided very interpretable models, we were able to achieve accuracy that matched that of the best black box model. At that point, we were puzzled about what to do. Should we play by the rules and provide a black box to the judges and try to explain it? Or should we provide the transparent, interpretable model? In other words, what do you do when you find you've been forced into the false dichotomy of the robot and the surgeon?

Our team decided that for a problem as important as credit scoring, we would not provide a black box to the judging team merely for the purpose of explaining it. Instead, we created an interpretable model that we thought even a banking customer with little mathematical background would be able to understand. The model was decomposable into different mini-models, where each one could be understood on its own. We also created an additional interactive online visualization tool for lenders and individuals. Playing with the credit history factors on our website would allow people to understand which factors were important for loan application decisions. No black box at all. We knew we probably would not win the competition that way, but there was a bigger point that we needed to make.

One might think there are a lot of applications where interpretable models cannot possibly be as accurate as black box models. After all, if you could build an accurate interpretable model, why would you then use a black box? However, as the Explainable Machine Learning Challenge revealed, there are actually a lot of applications where people do not try to construct an interpretable model, because they might believe that for a complex data set, an interpretable model could not possibly be as accurate as a black box. Or perhaps they want to preserve the model as proprietary. One might then consider that if interpretable deep-learning models can be constructed for computer vision and timeseries analysis (e.g., Chen et al., 2019; Y. Li et al., 2017; O. Li et al., 2018; Ming et al., 2019), then the standard should be changed from the assumption that interpretable models do *not* exist, to the assumption that they *do*, until proven otherwise.

Further, when scientists understand what they are doing when they build models, they can produce AI systems that are better able to serve the humans who rely upon them. In these cases, the so-called accuracy–interpretability tradeoff is revealed to be a fallacy: more interpretable models often become more (and not less) accurate.

The false dichotomy between the accurate black box and the not-so accurate transparent model has gone too far. When hundreds of leading scientists and financial company executives are misled by this dichotomy, imagine how the rest of the world might be fooled as well. The implications are profound: it affects the functioning of our criminal justice system, our financial systems, our healthcare systems, and many other areas. Let us insist that we do not use black box machine learning models for high-stakes decisions unless no interpretable model can be constructed that achieves the same level of accuracy. It is possible that an interpretable model can always be constructed—we just have not been trying. Perhaps if we did, we would never use black boxes for these high-stakes decisions at all.

Notes

1. The Explainable Machine Learning Challenge website is here: https://community.fico.com/s/explainable-machine-learning-challenge

2. This article is based on Rudin's experience competing in the 2018 Explainable Machine Learning Challenge.

3. Readers can play with our interactive competition entry for the challenge here: http://dukedatasciencefico.cs.duke.edu

4. Our entry indeed did not win the competition as judged by the competition's organizers. The judges were not permitted to interact with our model and its visualization tool at all; it was decided after the submission deadline that no interactive visualizations would be provided to the judges. However, FICO performed its own separate evaluation of the competition entries, and our entry scored well in their evaluation, earning the FICO Recognition Award for the competition. Here is FICO's announcement of the winners:

https://www.fico.com/en/newsroom/fico-announces-winners-of-inaugural-xml-challenge?utm_source=FICO-Community&utm_medium=xml-challenge-page

5. As far as the authors know, we were the only team to provide an interpretable model rather than a black box.

References

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research, 18*(234), 1-78.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Sydney, NSW, Australia, 721–1730.

Chen, C., Li, O., Barnett, A., Su, J., & Rudin, C. (2019). This looks like that: Deep learning for interpretable image recognition. Vancouver, Canada, *Advances in Neural Information Processing Systems*.

Li, O., Liu, H., Chen, C., & Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, Louisiana, 3530–3587.

Li, Y., Murias, M., Major, S., Dawson, G., Dzirasa, K., Carin, L., & Carlson, D. E. (2017). Targeting EEG/LFP synchrony with neural nets. *Advances in Neural Information Processing Systems*, Montreal, Canada, 4620–4630.

Ming, Y., Xu, P., Qu, H., & Ren, L. (2019). Interpretable and steerable sequence learning via prototypes. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, Alaska, 903–913.

Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. (2015). Population-level prediction of Type 2 Diabetes from claims data and analysis of risk factors. *Big Data, 3,* 277–287.

Angwin, J. and Larson, J. and Mattu, S. and Kirchner, L. Machine Bias. ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, accessed 2016-5-23.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1,* 206–215.

Rudin, C., & Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces, 48,* 449–466.

Rudin, C., Wang, C., & Coker, B. (2019). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review* (in press).

Tollenaar, N., & van der Heijden, P. G. M. (2013). Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 176*, 565–584.

Zeng, J., Ustun, B., & Rudin, C. (2016). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 180,* 689–722.