nature machine intelligence

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s42256-019-0048-x

In the format provided by the authors and unedited.

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Duke University, Durham, NC, USA. e-mail: cynthia@cs.duke.edu

Supplementary Materials for ‘Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead’

Cynthia Rudin
Duke University
cynthia@cs.duke.edu

A On the Two Types of Black Box

Black box models of the first type are too complicated for a human to comprehend, and black box models of the second type are proprietary. Some models are of both types. The consequences of these two types of black box are different, but related. For instance, for a black box model that is complicated but not proprietary, we at least know what variables it uses. We also know the model form and could use that to attempt to analyze the different parts of the model. For a black box model that is proprietary but not complicated [we have evidence that COMPAS is such a model, see Rudin et al., 2018], we may not even have access to query it in order to study it. If a proprietary model is too sparse, there is a risk that it could be easily reverse-engineered, thus there is an incentive to make proprietary models complicated in order to preserve their secrecy.

B Performance Comparisons

For most problems with meaningful structured covariates, machine learning algorithms tend to perform similarly, with no algorithm clearly dominating the others. The variation due to tuning parameters of a single algorithm can often be higher than the variation between algorithms. This lack of single dominating algorithm for structured data is arguably why the field of machine learning focuses on applications like image and speech recognition, whose data are represented in raw features (pixels, sound files); these are fields for which the choice of algorithm impacts performance. Even for complex domains such as medical records, it has been reported in some studies that logistic regression has identical performance to deep neural networks [e.g. Razavian et al., 2015].

If there is no dominating algorithm, the Rashomon Set argument discussed in the main text would suggest that interpretable models might perform well.

Unfortunately the culture of publication within machine learning favors selective reporting of algorithms on selectively chosen datasets. Papers are often rejected if small or no performance gains are reported between algorithms. This encourages omission of accurate baselines for comparison, as well as omission of datasets on which the method does not perform well, and encourages authors to poorly tune the parameters of baseline methods, or equivalently, place more effort into tuning the parameters of the author’s own method. This creates an illusion of large performance differences between algorithms, even when such performance differences do not truly exist.

C Counterfactual Explanations

Some have argued that counterfactual explanations [e.g., see Wachter et al., 2018] are a way for black boxes to provide useful information while preserving secrecy of the global model. Counterfactual explanations, also called inverse classification, state a change in features that is sufficient (but not necessary) for the prediction to switch to another class (e.g., “If you reduced your debt by $5000 and increased your savings by $50% then you would have qualified for the loan you applied for”). In some situations, recourse is more important than a standard counterfactual explanation; recourse is a special type of counterfactual explanation where the user is able to take an action to reverse a decision [Ustun et al., 2019]. Here, the users must realistically be able to adjust the features to change the decision (the counterfactual explanation cannot be “to change the decision you must become five years younger" for instance).

There are several problems with the argument that counterfactual explanations are sufficient. For loan applications, for instance, we would want the counterfactual explanation to provide the *lowest cost* action for the user to take, *according to the user's own cost metric*. [See Chang et al., 2012, for an example of lowest-cost counterfactual reasoning in product rankings]. In other words, let us say that there is more than one counterfactual explanation available (e.g., the first explanation is "If you reduced your debt by $5000 and increased your savings by $50% then you would have qualified for the loan you applied for" and the second explanation is "If you had gotten a job that pays $500 more per week, then you would have qualified for the loan"). In that case, the explanation shown to the user should be the easiest one for the user to actually accomplish. However, it is unclear in advance which explanation would be easier for the user to accomplish. In the credit example, perhaps it is easier for the user to save money rather than get a job or vice versa. In order to determine which explanation is the lowest cost for the user, we would need to elicit cost information for the user, and that cost information is generally very difficult to obtain; worse, the cost information could actually change as the user attempts to follow the policy provided by the counterfactual explanation (e.g., it turns out to be harder than the user thought to get a salary increase). For that reason it is unclear that counterfactual explanations would suffice alone for high stakes decisions; interpretable models paired with recourse options would be better. Additionally, counterfactual explanations of black boxes have many of the other pitfalls discussed throughout this paper.

D Interpretable Models that Provide Smaller-Than-Global Explanations

It is possible to create a global model (perhaps a complicated one) for which explanations for any given individual are very sparse. In other words, even if the global model would take several pages of text to write, the prediction for a given individual can be very simple to calculate (perhaps requiring only 1-2 conditions). Let us consider the case of credit risk prediction. Assume we do not need to justify to the client why we would grant a loan, but we would need to justify why we would deny a loan.

Let us consider a disjunctive normal form model, which is a collection of "or's" of "and's." For instance, the model might deny a loan if "(credit history too short AND at least one bad past trade) OR (at least 4 bad past trades) OR (at least one recent delinquency AND high percentage of delinquent trades)." Even if we had hundreds of conjunctions within the model, only one of these needs to be shown to the client; if any conjunction is true, that conjunction is a defining reason why the client would be denied a loan. In other words, if the client had "at least one recent delinquency AND high percentage of delinquent trades," then regardless of any other aspects of her credit history, she could be shown that simple explanation, and it would be a defining reason why her loan application would be denied.

Disjunctive normal form models are well-studied, and are called by various names, such as "or's of and's," as well as "decision rules," "rule sets" and "associative classifiers." There has been substantial work in being able to generate such models over the past few years so that the models are globally interpretable, not just locally interpretable (meaning that the global model consists of a small number of conjunctions) [e.g., see Dash et al., 2018; Goh and Rudin, 2014; Rijnbeek and Kors, 2010; Su et al., 2016; Wang et al., 2017].

There are many other types of models that would provide smaller-than-global explanations. For instance, falling rule lists [Chen and Rudin, 2018; Wang and Rudin, 2015] provide shorter explanations for the decisions that are most important. For instance, a falling rule list for predicting patient mortality would use few logical conditions to categorize whether a patient is in a high-risk group, but use several additional logical conditions to determine which low-risk group a patient falls into.

E Algorithm Stability

A common criticism of decision trees is that they are not stable, meaning that small changes in the training data lead to completely different trees, giving no guidance as to which tree to choose. In fact, the same problem can happen in *linear* models when there are highly correlated features. This can happen even in basic least squares, where correlations between features can lead to very different models having precisely the same levels of performance. When there are correlated features, the lack of stability happens with most algorithms that are not strongly regularized.

I hypothesize this instability in the learning algorithm could be a side-effect of the Rashomon effect mentioned in the main text – that there are many different almost-equally good predictive models. Adding regularization to an algorithm increases stability, but also limits flexibility of the user to choose which element of the Rashomon set would be more desirable.

For applications where the models are purely predictive and not causal (e.g., in criminal recidivism where we use age and prior criminal history to predict future crime), there is no assumption that the model represents how outcomes are actually generated. The importance of the variables in the model does not reflect a causal relationship between the variables and the outcomes. Thus, without additional guidance from the domain expert, there is no way to proceed further to choose a single "best model" among the set of models that perform similarly. As discussed above, regularization can act as this additional input.

I view the lack of algorithmic stability as an advantage rather than a disadvantage. If the lack of stability is indeed caused by a large Rashomon effect, it means that domain experts can add more constraints to the model to customize it without losing accuracy.

In other words, while many people criticize methods such as decision trees for not being stable, I view that as a strength of interpretability for decision trees. If there are many equally accurate trees, the domain expert can pick the one that is the most interpretable.

Note that not all researchers working in interpretability agree with this general sentiment about the advantages of instability [Murdoch et al., 2019].

References

A. Chang, C. Rudin, M. Cavaretta, R. Thomas, and G. Chou. How to reverse-engineer quality rankings. *Machine Learning*, 88:369–398, September 2012.

C. Chen and C. Rudin. An optimization approach to learning falling rule lists. In *Proceedings of Machine Learning Research Vol. 84: Artificial Intelligence and Statistics (AISTATS)*, pages 604–612, 2018.

S. Dash, O. Günlük, and D. Wei. Boolean decision rules via column generation. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

S. T. Goh and C. Rudin. Box drawings for learning with imbalanced data. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.

W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv e-prints: 1901. 04592* [*statistical machine learning*], Jan. 2019.

N. Razavian et al. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, 3 (4), 2015.

P. R. Rijnbeek and J. A. Kors. Finding a short and accurate decision rule in disjunctive normal form by exhaustive search. *Machine Learning*, 80(1):33–62, July 2010.

C. Rudin, C. Wang, and B. Coker. The age of secrecy and unfairness in recidivism prediction. *arXiv e-prints 1811. 00731* [*applied statistics*], Nov. 2018.

G. Su, D. Wei, K. R. Varshney, and D. M. Malioutov. Interpretable two-level boolean rule learning for classification. In *Proceedings of ICML Workshop on Human Interpretability in Machine Learning*, pages 66–70, 2016.

B. Ustun, A. Spangher, and Y. Liu. Actionable Recourse in Linear Classification. In *ACM Conference on Fairness, Accountability and Transparency (FAT*)*, 2019.

S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 1(2), 2018.

F. Wang and C. Rudin. Falling rule lists. In *Proceedings of Machine Learning Research Vol. 38: Artificial Intelligence and Statistics (AISTATS)*, pages 1013–1022, 2015.

T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille. A bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, 18(70):1–37, 2017.