Assessing the Local Interpretability of Machine Learning Models*

Dylan Slack [†] **Sorelle Friedler** [‡] **Carlos Scheidegger** [§] **Chitradeep Dutta Roy** [¶]

Abstract

This work defines a proxy metric for limited notions of interpretability and makes basic comparisons across model classes using the metric. We focus on two definitions of interpretability that have been introduced in the machine learning literature: simulatability and "what if" local explainability . Through a user study with 1000 participants, we test whether humans perform well on tasks that mimic the definitions of simulatability and "what if" local explainability on models that are typically considered locally interpretable. We propose a metric - the runtime operation count on the simulatability task - to indicate the relative interpretability of models and show that as the number of operations increases the users' accuracy on the local interpretability tasks decreases. We find evidence consistent with the common intuition that decision trees and logistic regression models are interpretable and are more interpretable than comparable neural networks according to our proposed metric.

1 Introduction

While many interpretable methods have been considered (e.g., see surveys [1, 2]) there has been comparatively little work on assessing whether the definitions of interpretability make sense from a human-performance perspective [3, 4]. Perhaps most related, Lage. et. al consider simulation, verification, and counter factual reasoning for decision sets [5]. The problem of measuring human interpretability of models is complex, including cognitive, representational, model-choice, and context-dependent aspects. This work focuses on measuring the impact of model-choice on user interpretability, purposefully leaving the other aspects that contribute to interpretability fixed so that future studies can consider the effects of varying those variables. Within that limited scope, our goal is to determine the relative "interpretability" of model types. Interpretability can be broadly divided into *global interpretability*, meaning understanding the entirety of a trained model including all decision paths, and *local interpretability*, the goal of understanding the results of a trained model on a specific input and small deviations from that input. In this paper, we focus on local interpretability, and on two specific definitions.

We perform a user study to assess *simulatability* [6] - here interpreted as the ability of a person to run a model and get the correct output (model classification) for a given input - and *"what if" local explainability* [7, 6] - information that helps a user determine how small changes to a given input

*This research was funded in part by the NSF under grant IIS-1633387.

[†]University of California Irvine, Department of Computer Science, work done while an undergraduate at Haverford College

[‡]Haverford College, Department of Computer Science

[§]University of Arizona, Department of Computer Science

[¶]University of Utah, Deartment of Computer Science

Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. (non-archival)

affect the model classification.[6] We will refer to a model as *locally interpretable* if users are able to correctly perform *both* of these tasks when given a model and input.

The main contributions of this work are to (1) substantiate the folk hypothesis that decision trees and logistic regressions are more locally interpretable than neural networks and (2) provide evidence that the total run time operation count needed to classify an input can be useful as a metric for the local interpretability of a machine learning model.

2 A Metric for Local Interpretability

Motivated by the previous literature and its calls for user-validated metrics that capture aspects of interpretability [8], we wish to assess whether a candidate metric captures a user's ability to simulate *and* "what if" locally explain a model. The candidate metric we consider here is the *total number of runtime operation counts* performed by the model to determine the classification of a given input. Effectively, we seek a proxy for the work that a user must do (in their head or via a calculator) in order to simulate a model on a given input, and will claim that the total number of operations also impacts a user's ability to perform a "what if" local explanation of a model. If true, this second claim is more surprising, since understanding how local pertubations to an input result in changed outputs *without* rerunning the model is a more complex task. In order to calculate the number of runtime operations for a given input, we instrumented the prediction operation for existing trained models in python's scikit-learn package.[7]

3 User Study

We designed a crowdsourced experiment that was given to 1000 participants. Participants were asked to run a model on a given input and then evaluate the same model on a locally changed version of the input. For this study we consider the local interpretability of three models: decision trees, logistic regression, and neural networks. The models were trained using the standard package scikit-learn. Training details are given in the appendix in section 6.1.

Our decision tree representation is a standard node-link diagram representation for a decision tree or flow chart. In order to allow users to simulate the logistic regression and neural network classifiers we needed a representation that would walk the users through the calculations without previous training in using the model or any assumed mathematical knowledge beyond arithmetic. We created a "fill in the blank" style logistic regression representation that walked users through the process of solving a logistic regression. The resulting representation for logistic regression is shown in the appendix in Figure 2. The neural network representation used the same representation as the logistic regression for each node and one page per layer. In order to allow users to assess the "what if" local explainability of the model, we also asked them to determine the output of the model for a perturbed version of the initial input they were shown. In the perturbed setting, the user's answers for the unperturbed problem are shown, but the user is *not* led back through the "fill in the blank" exercise.

In order to avoid effects from study participants with domain knowledge, we created synthetic datasets to train the models. We created four synthetic datasets simple enough so that each model could achieve 100% test accuracy. These four datasets were used to train the three considered models via an 80/20 train-test split. We generated user inputs using the test data. For each test data point, we changed one dimension incrementally in order to create a perturbed input.

We used Prolific to distribute the survey to 1000 users each of whom was paid $3.50 for completing it. Participants were restricted to those with at least a high school education (due to the mathematical nature of the task) and a Prolific rating greater than 75 out of 100. The full survey information (hosted through Qualtrics), preregistered hypotheses, and resulting data is available online.[8] Further study details can be found in the appendix.

[6] This is similar to notions sometimes referred to as counterfactual explanations.

[7] `https://github.com/darkreactions/measuring_interpretability/`

[8] See footnote 6.

Table 1: Comparative correct distributions and p-values between model types generated through Fisher Exact Tests for confident responses. Relative correctness is shown for simulatability (correctness on the original input (Sim.)), "what if" local explainability (correctness on the perturbed input (What If)), and local interpretability (correctness on both parts). Decision trees (DT), logistic regression (LR), and neural networks (NN) are considered.

ALT. HYPO.		SIM.	WHAT IF	LOCAL INTERP. (BOTH)
DT > NN	DT CORR.	717 / 930	719 / 930	594 / 930
	NN CORR.	556 / 930	499 / 930	337 / 930
	P-VALUE	1.5×10^{-14}	7.3×10^{-26}	9.3×10^{-32}
	95% CI	$[1.69, \infty]$	$[2.20, \infty]$	$[2.36, \infty]$
DT > LR	DT CORR.	717 / 930	719 / 930	594 / 930
	LR CORR.	592 / 930	579 / 930	425 / 930
	P-VALUE	3.7×10^{-9}	2.6×10^{-11}	5.9×10^{-14}
	95% CI	$[1.43, \infty]$	$[1.54, \infty]$	$[1.60, \infty]$
LR > NN	LR CORR.	592 / 930	579 / 930	425 / 930
	NN CORR.	556 / 930	499 / 930	337 / 930
	P-VALUE	1.3	2.9×10^{-3}	5.7×10^{-4}
	95% CI	$[0.90, \infty]$	$[1.09, \infty]$	$[1.13, \infty]$

4 User Study Results

In order to assess the local interpretability of different model types, we first separately consider the user success on the task for simulatability (the original input) and the task for "what if" local explainability (the perturbed input). In the study inputs were chosen so that 50% of the correct model outputs were "yes" and 50% were "no". Thus, we compare the resulting participant correctness rates to the null hypothesis that respondents are correct 50% of the time using and exact binomial test.

The results given in table 2 in the appendix indicate strong support for the simulatability of decision trees, logistic regression, and neural networks based on the representations the users were given. The results also indicate strong support for the "what if" local explainability of decision trees and logistic regression models, but neural networks were *not* found to be "what if" locally explainable. We thus have evidence that suggests decision trees and logistic regression models are locally interpretable and neural networks are not.

In order to assess the relative local interpretability of models — to evaluate the folk hypothesis that decision trees and logistic regression models are more interpretable than neural networks — we compared the distributions of correct and incorrect answers on both tasks across pairs of model types. We applied one-sided Fisher exact tests with the null hypothesis that the models were equally simulatable, "what if" locally explainable, or locally interpretable.

The results, presented in Table 1, give strong evidence that decision trees are more locally interpretable than logistic regression or neural network models on both the simulatability and "what if" local explainability tasks in terms of operation count. Interestingly, while there was strong evidence that logistic regression is more "what if" locally explainable and more locally interpretable than neural networks, there is not evidence that logistic regression models are more simulatable than neural networks using the given representations. This may be because the logistic regression and neural network representations were very similar. An analysis of the users who got both tasks right, i.e., were able to locally interpret the model, shows that the alternative hypothesis was strongly supported in all three cases, thus supporting the folk hypotheses that decision trees and logistic regression models are more interpretable than neural networks.

We considered the relationship between total operation counts, time, and accuracy on the simulatability, "what if" local explainability, and combined local interpretability tasks. The graphs showing these relationships, including ellipses that depict the degree to which the different measurements are linearly related to each other, are shown in Figure 1. Across all three interpretability tasks it appears clear that as the number of operations increases, the total time taken by the user also increases (see the first row of Figure 1). This effect is perhaps not surprising, since the operation count considered is for the simulatability task and the representations given focus on performing each operation. Perhaps

more surprisingly, as the total operation count on the simulatability task increases, the total time taken on the "what if" local explainability task also increases; though that pattern is most clear for the decision tree models. When considering the combined local interpretability task, this upward trend in time is also apparent.

In the second row of Figure 1, we can see that, as the total number of runtime operations increases, the accuracy decreases for all three interpretability tasks for the decision tree models, but there is no clear trend for the logistic regression and neural network models. This lack of effect may be due to the comparatively smaller range of operation counts examined for these two model types, or it may be that the local interpretability of these model types is not as related to operation count as it is for decision trees. The lack of overlap in the ranges for the operation counts of logistic regression and neural networks also makes it hard to separate the effects of the model type on the results.

5 Discussion and Conclusion

We investigated the local interpretability for decision trees, logistic regressions, and neural networks and showed support via a user study for the folk hypotheses that decision trees and logistic regression models are locally interpretable while neural networks are not. We introduced the run time operation count local interpretability metric and showed that the number of runtime operations has a positive relationship to the time a user takes to locally interpret a model and a negative relationship to the users' accuracy on the local interpretation task (the ability to both simulate and "what if" locally explain a model). The introduction of this metric opens the possibility of analyzing other model types for their local interpretability without running a user study.

References

[1] Christoph Molnar. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/, 2018. `https://christophm.github.io/interpretable-ml-book/`.

[2] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):93, 2018.

[3] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. Transparent and Interpretable Machine Learning in Safety Critical Environments Workshop at NIPS, 2017.

[4] Hiva Allahyari and Niklas Lavesson. User-oriented assessment of classification model understandability. In *11th scandinavian conference on Artificial intelligence*. IOS Press, 2011.

[5] Isaac Lage, E. Chen, J. He, M. Narayanan, S. Gershman, B. Kim, and F. Doshi-Velez. An evaluation of the human-interpretability of explanation. 2018.

[6] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):30, 2018.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

[8] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

6 Appendix

6.1 Additional Training Details

Decision trees were trained using `sklearn.tree.DecisionTreeClassifier` without any depth restrictions and with default parameters. Logistic regression was trained using

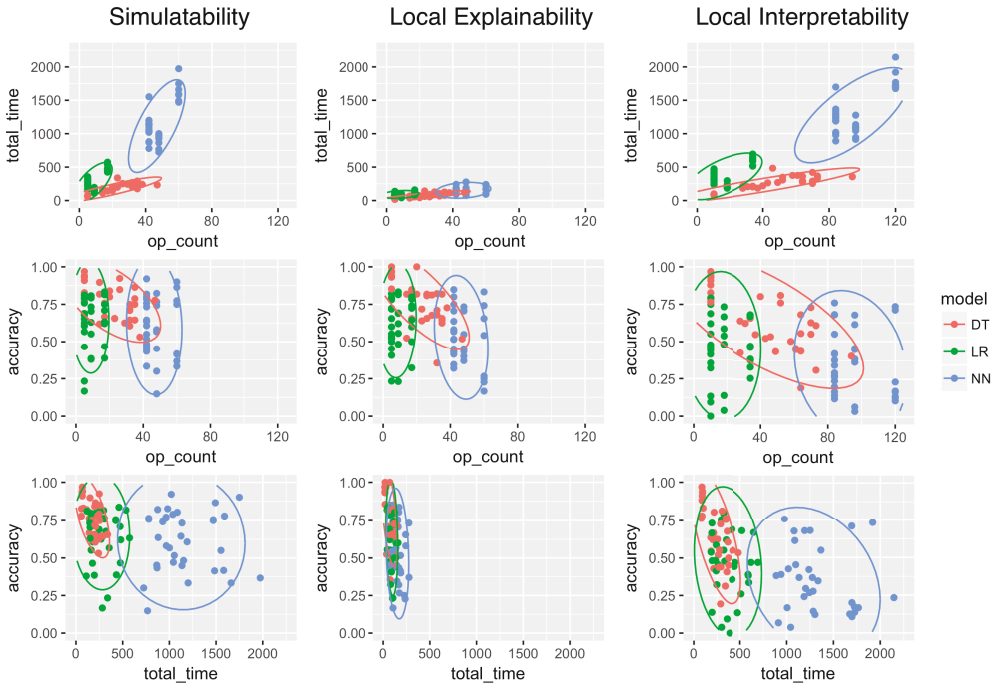

Figure 1: Comparisons shown are between total operations for a particular trained model and input, the time taken by the user to complete the task, and the accuracy of the users on that task for the simulatability (original input), "what if" local explainability (perturbed input), and the combined local interpretability (getting both tasks correct) tasks. The total time shown is in seconds. The total operation count is for the simulatability task on the specific input; this is the same for both "what if" local explainability and simulatability except for in the case of the decision tree models, where operation counts differ based on input. The local interpretability operation count is the sum for the simulatability and "what if" local explainability task operation counts. Accuracy shown is averaged over all users who were given the same input for that task and trained model. The models considered are decision trees (DT), logistic regression models (LR), and neural networks (NN). The ellipses surrounding each group depict the covariance between the two displayed variables, and capture 95% of the sample variance.

`sklearn.linear_model.LogisticRegression` with the multi_class argument set to '*multinomial*' and '*sag*'(Stochastic average gradient descent) as the solver. The neural network was implemented using `sklearn.neural_network.MLPClassifier`. The neural network used is a fully connected network with 1 input layer, 1 hidden layer with 3 nodes, and 1 output layer. The `relu` (rectified linear unit) activation function was used for the hidden layer.

6.2 Additional Study Details

Preregistered Hypotheses We preregistered two experimental hypotheses. Namely, that time to complete will be positively related to operation count and that accuracy will be negatively related to operation count. We also preregistered two exploratory hypotheses. These were that we would explore the specific relationship between time and accuracy versus operation count and that we would explore how the perturbed input is related to time and operation count. These hypotheses can be found at the Open Science Framework at: *url removed for anonymization*

Study Setup Issues After running the user study, we found that an error in the survey setup meant that the survey exited prematurely for users given two of the eight inputs on the decision tree models for one dataset. Since we did not receive data from these participants, Prolific recruited other participants who were allocated to other inputs and datasets, so the analyzed dataset does not include data for these two inputs. Users who contacted us to let us know about the problem were still paid.

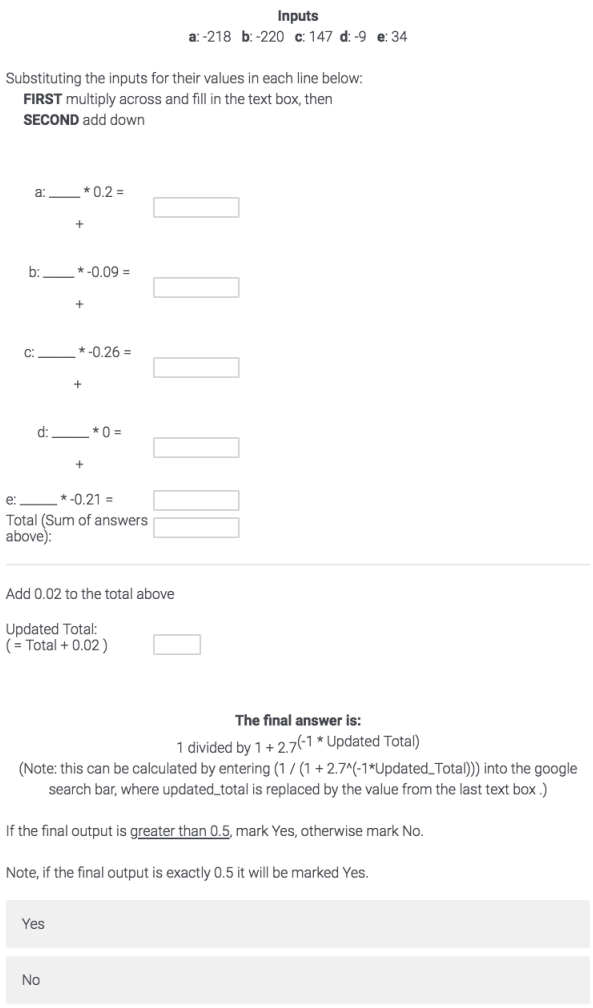

Figure 2: The logistic regression representation shown to users.

Multiple Comparison Corrections In order to mitigate the problem of multiple comparisons, all p-values and confidence intervals we report in the next section include a Bonferroni correction factor of 28. While we include 15 statistical tests in this paper, we considered a total of 28. Reported p-values greater than one arise from these corrections.

Table 2: Per-model correct responses out of the total confident respondents on the original input (simulatability task (Sim)) and perturbed inputs ("what if" local explainability task (What If)) for decision trees, logistic regression, and neural networks. p-values given are with respect to the null hypothesis that respondents are correct 50% of the time, using exact binomial tests.

Model Type		Sim.	What If
Decision Tree	Correct	717 / 930	719 / 930
	p-Value	5.9×10^{-63}	5.16×10^{-64}
	95% CI	$[0.73, 0.81]$	$[0.73, 0.82]$
Logistic Regression	Correct	592 / 930	579 / 930
	p-Value	1.94×10^{-15}	2.07×10^{-12}
	95% CI	$[0.59, 0.69]$	$[0.57, 0.67]$
Neural Network	Correct	556 / 930	499 / 930
	p-Value	7.34×5.5^{-8}	0.78
	95% CI	$[0.55, 0.65]$	$[0.49, 0.59]$