Evaluating Perturbation-based Post-hoc Explainer Fidelity Using Inherently Interpretable Models

Adam Sun
Stanford University, Wells Fargo
adsun@stanford.edu

Agus Sudjianto
Wells Fargo
Agus.Sudjianto@wellsfargo.com

Aijun Zhang
Wells Fargo
Aijun.Zhang@wellsfargo.com

Abstract

Post-hoc explainability methods like LIME and SHAP are frequently used to explain predictions of complex black-box machine learning models like neural networks and model ensembles. However, these explainers are only approximations, which is an issue in high-risk fields like finance, medicine, and policy, where more exact interpretations are preferred and necessary. This work aims to gauge the fidelity, or faithfulness, that LIME and SHAP have to the internal decision processes of the model. To do so, we utilize inherently interpretable models to obtain exact interpretations, comparing them to the approximations generated by LIME and SHAP. We experiment with a wide variety of different real-world datasets, models, and model complexities to finally assess the fidelity of LIME and SHAP.

This work shows that LIME and SHAP do not provide reliable and consistently faithful explanations. We demonstrate that LIME and SHAP's fidelity varies widely between datasets and models, and rapidly decreases with increasing model complexity. This work further reinforces the need to use inherently interpretable models that provide exact and consistent interpretations that can be relied upon.

1. Introduction

Machine learning (ML) is a constantly developing field that is being increasingly used in a wide variety of fields. However, many of the most commonly used algorithms in machine learning – from Support Vector Machines (SVM) to XGBoost to deep neural networks (DNNs) – are black-box models. These models can contain millions of parameters portraying very complex patterns, causing them to be too complicated to be easily understood by humans. This is a big problem in high-risk areas, including in finance, medicine, and law, where a model's decisions can make significant impacts on the lives of human beings.

A very common approach to combat this issue is to gauge local interpretability, where feature importance, or the importance of each feature on a prediction, is gauged based on an individual point. This allows individual predictions to be explained and further understood by researchers.

One approach to local interpretability is perturbation-based model agnostic post-hoc explainability methods like LIME and SHAP, where a second model is utilized to explain a trained model's individual predictions. These methods estimate the contribution of each feature to the output by perturbing the input and observing the model's response to the perturbations [18]. With the coefficients/SHAP values from these explainers, researchers can obtain local explanations of individual model predictions. While these methods can work for any model, they are mostly used for black box models like neural networks.

An alternative approach to post-hoc explainability methods is using inherently interpretable models, which are models designed in a manner so that exact local interpretations of the model can be obtained. Examples of inherently interpretable models include linear regression models, GLM, GAM, Decision Trees, EBM [13], and GAMI-Net [23]. Unlike post-hoc explainability methods, using interpretations of these models guarantees a faithful understanding of the model.

Despite the fact that using inherently interpretable models grants an exact understanding of the model, most researchers decide to utilize black box models and explain them with post-hoc explainability methods. This is due to the commonly-held belief that black box models have greater predictive power, can be more accurate, and can learn "hidden patterns" in data [16] [15]. However, inher-

ently interpretable models have been shown to perform just as well as black-box models on high risk tasks, while still remaining completely transparent [4] [3].

Fidelity is an explainer's faithfulness to the original model. There are two types of fidelity – external and internal fidelity [11]. The former concerns fidelity to model predictions, while the latter concerns fidelity to the internal decision processes of a model. Since explanations are by default approximations, they will not be internally faithful to the original model, making it a bad idea to rely on them for high-risk applications [15]. Indeed, using single points to attempt to understand a complex is too optimistic and naive, and may not reflect the model as a whole [2]. Using unfaithful explanations to explain a model can be very misleading, causing researchers and users alike to develop false understanding surrounding a model, potentially resulting in unethical and costly consequences.

In this paper, we prove the infidelity of perturbation-based post-hoc explainability methods by directly investigating and measuring the fidelity of LIME and SHAP on real-world datasets, in fields ranging from medicine and biology to physics and finance. We experiment with models of differing complexities, and also compare local feature importance interpretations from a range of inherently interpretable models with post-hoc explanations on the same models to measure the faithfulness of these methods. Our work shows that the fidelity of LIME and SHAP to the ground truth of inherently interpretable models varies widely.

Model development in this paper makes extensive use of the PiML toolbox [21], an interactive Python toolbox for interpretable machine learning model development and validation.

2. Background

2.1. Post-hoc Explainability Methods

The most commonly used perturbation-based post-hoc explainability methods are LIME and SHAP. These methods solely rely on model inputs and outputs to make inferences about the model itself.

2.1.1 Local Interpretable Model-Agnostic Explanations (LIME) [14]

LIME utilizes an interpretable local surrogate model to approximate the particular example's neighborhood. To do so, perturbations are made surrounding the point of interest, and black box predictions are made on each of the points. Each perturbation is weighted on the distance to the point, and an interpretable model is trained based on the perturbed dataset. By doing so, the interpretable model can then be used to explain the original prediction.

For LIME, the explanation for instance x are expressed as

$$explanation(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g) \quad (1)$$

where f is the original model that is to be explained, g is the interpretable model (usually linear) used to explain the prediction, L is the loss. Each pi_x is the proximity measure used to set the size of the kernel used to weight the perturbations, and $\Omega(g)$ is the model complexity used to penalize g [14].

2.1.2 SHapley Additive exPlanations (SHAP) [10]

SHAP (SHapley Additive exPlanations) computes the contributions of each feature to the model, creating a linear additive feature attribution model.

To do so, SHAP utilizes Shapley values from cooperative game theory, which are computed as

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (2)$$

where S is a subset of F, all features in the model. and $f_{S \cup \{i\}}$ and f_S are models trained with and without feature i present. [10]

KernelSHAP (SHAP's model-agnostic method) utilizes a regression-based method based on LIME to estimate Shapley values, and specifies the explanation as

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \quad (3)$$

where g is the explanation model, M is the maximum coalition size, and $z' \in \{0, 1\}^M$ is a coalition vector used to specify which features are being used in the coalition. [10] [12].

2.2. Interpretable Models

In order to fully gauge the fidelity of LIME and SHAP, we need to utilize models with a wide range of different decision boundaries. So, we use the following interpretable models:

2.2.1 ReLU-DNN with Aletheia Unwrapper

Deep neural networks with the Rectified Linear Unit activation function (ReLU-DNNs) are normally black boxes. However, they express a piecewise linear function and partition the input space into a finite set of convex activation regions, each with its own activation pattern.

The Aletheia toolbox includes an unwrapper that can unwrap a ReLU-DNN into an equivalent set of local linear

models based on these activation patterns. Each local linear model functions exclusively on a disjoint convex sub-region of the input space. The weights of these local linear models can then be used to exactly interpret the ReLU-DNN locally, obtaining the exact partial dependence on each feature [19].

Additionally by imposing a sparsity constraint through ℓ_1 regularization, the number of LLMs making up the ReLU-DNN can be reduced, thus reducing the complexity of the model [20]. These simplified models can be easier to interpret.

3. Related Work

There has been a consensus surrounding the lack of fidelity of post-hoc explainability methods, especially after Cynthia Rudin's paper highlighting the dangers in relying upon them for high-stakes decisions [15]. As a result, there have been numerous works aiming to assess fidelity of a model. These methods can be categorized into dataset-based methods and perturbation-based methods.

The first approach to assess fidelity is to construct synthetic datasets which allow for efficient computation of conditional expected values that can be used to evaluate explanations [8]. However, this approach is problematic since post-hoc explainers are intended to explain the model, not the data itself. So, one approach to address the issues with synthetic data generation is SynthGauss [1]. This approach generates synthetic datasets consisting of clusters with points sampled from Gaussian distributions. Nevertheless, synthetic data is not necessarily be reflective of real-world datasets.

A second method to assess model fidelity are perturbation-based methods. These methods are used when there is no ground truth available. Examples of such methods include Prediction Gap on Important Feature Perturbation (PGI) and Prediction Gap on Unimportant Feature Perturbation (PGU) [1], measuring the difference in prediction accuracy after perturbing important features. Alternatively, the $Top_jSimilarity$ metric compares the SHAP values of the original model with the surrogate explanation model [11]. RemOve And Retrain (ROAR) removes features deemed important to each explainer and retrains the model to gauge fidelity [6]. However, these methods do not gauge the explainers' internal fidelity, since they are based on external perturbation to model inputs.

Obtaining a true measure of fidelity involving internal fidelity can be problematic, especially when there is no exact ground truth feature ranking available to the model. So, an ideal approach is to utilize an exact local interpretation to use as a ground truth for comparison of explainability methods.

4. Methods

4.1. Approach

Our approach to assess the fidelity of explainability methods by utilizing the Aletheia toolbox and the exact local interpretations that it provides.

LIME, SHAP and our exact interpretations all assign feature importance values to each individual feature, measuring the contribution it has on a prediction. Ranking these values from largest to smallest magnitude creates feature importance rankings that can be compared [7]. Utilizing exact feature importance rankings from inherent interpretations insures that we are assessing exact internal fidelity, or how faithful LIME and SHAP are to the exact decision boundaries of the model. Additionally, this allows us to utilize real-world datasets instead of synthetic ones. We compare feature importance rankings from LIME, KernelSHAP, and exact interpretations, measuring the agreement of these rankings.

We use default author implementations for LIME and KernelSHAP, leaving heuristic definition of hyperparameters for future work.

4.2. Metrics

To measure the agreement between the feature importance rankings from the ground truth and each explanation, we utilize both top-k metrics and ranking metrics. A higher value of these metrics means a higher similarity for the corresponding explanation method.

4.2.1 Top-k metrics

Top-k metrics are used to measure the agreement for the k most important features. We take these metrics from Krishna et. al [7]. See Figure 1.

Feature Agreement measures the fraction of common features between the top-k most important features in two rankings. Rank Agreement, Sign Agreement, and Signed Rank Agreement measure the fraction of common features in the top-k features that have the same position, sign, and both position and sign, respectively, between the two rankings. As such, Signed Rank Agreement is the most strict metric.

In our experiments, we use $k = 5$.

$$FeatureAgreement(E_a, E_b, k) = \frac{|TopFeatures(E_a, k) \cap TopFeatures(E_b, k)|}{k}$$

$$RankAgreement(E_a, E_b, k) = \frac{|\bigcup_{s \in S} s \in TopFeatures(E_a, k) \wedge s \in TopFeatures(E_b, k) \wedge rank(E_a, s) = rank(E_b, s)|}{k}$$

$$SignAgreement(E_a, E_b, k) = \frac{|\bigcup_{s \in S} s \in TopFeatures(E_a, k) \wedge s \in TopFeatures(E_b, k) \wedge sign(E_a, s) = sign(E_b, s)|}{k}$$

$$SignedRankAgreement(E_a, E_b, k) = \frac{|\bigcup_{s \in S} s \in TopFeatures(E_a,k) \wedge s \in TopFeatures(E_b,k) \wedge rank(E_a,s) = rank(E_b,s) \wedge sign(E_a,s) = sign(E_b,s)|}{k}$$

Figure 1. Top-k metric calculations for explanations E_a and E_b. S denotes the set of all features. $TopFeatures(E, k)$ denotes the k top features of E, $rank(E, s)$ denotes the rank of feature s in E, and $sign(E, s)$ denotes the sign of feature s in E [7].

4.2.2 Ranking Correlation Metrics

Ranking correlation metrics can be used to compare the relative ordering of two different feature rankings. However, it is worth noting that the ordering of the highest-ranking features is very important information when using explanations to understand a model [7]. So, an ideal correlation metric should weigh higher ranked features more heavily than lower ranked ones.

So, we utilize a weighted version of Kendall's Tau [17] and Rank-Biased Overlap (RBO) [22]. These weighted metrics are able to provide a measure of the similarity of two different rankings, while giving higher-ranked inconsistencies a higher weight than lower-ranked ones. This allows for a realistic assessment of ranking agreement that can be directly applicable to human interpretation of explanations.

5. Experiments

5.1. Unwrapped ReLU-DNNs

Our first experiment explores LIME and SHAP's fidelity on ReLU deep neural networks with varying complexities. Neural networks are a very common state-of-the-art approach to modelling complex data. However, with the Aletheia unwrapper, we can use ReLU-DNNs as glass-box models and obtain local interpretations that we can compare to explanations. [19].

We use ReLU-DNNs that consist of 2 layers of 20 nodes each. We apply ℓ_1 regularizations of varying magnitudes to regularize the model and control the number of LLMs, resulting in a set of ReLU-DNNs with varying complexities. We then train these ReLU-DNNs on the German Credit dataset [5] for 20 epochs using an Adam optimizer with a learning rate of 0.001 and a batch size of 256. We sample a random set of 50 samples from the test set.

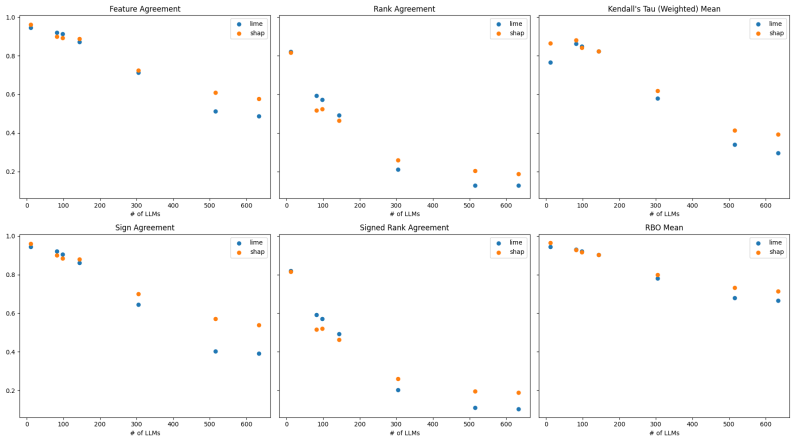
As the number of LLMs increases, the model begins to overfit, as seen in Figure 3. This is because models that are too complex tend to learn the training data too closely, including unavoidable noise [24]. Overfitting then causes test AUC to rapidly fall and train AUC to rapidly increase. Typically, ReLU-DNNs with a large number of LLMs have complex decision boundaries that do not generalize well to the test data.

Next, using the exact local feature importance rankings obtained from each ReLU-DNN, we calculate similarity metrics from Section 4.2 to gauge the fidelity of LIME and SHAP on models of varying complexity, then average the metrics across the 50 examples.

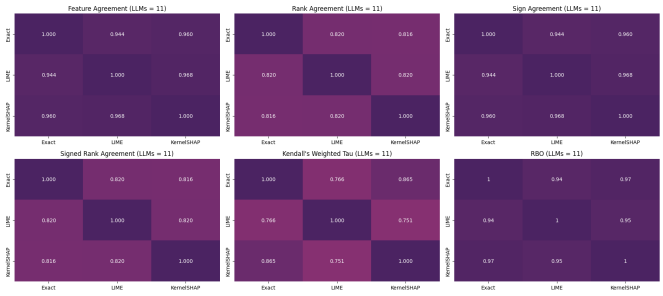
Our results indicate that model fidelity decreases monotonically at increasing LLM counts (see Figure 2). As the number of LLMs increases and the decision boundaries become more nonlinear, the approximations given by LIME and SHAP become less and less consistent with exact results, on average. In fact, at 634 LLMs, LIME and SHAP have an average rank agreement of only 0.128 and 0.188, respectively, demonstrating that at most 1 of the top 5 features of LIME and SHAP agree with the ground truth in terms of rank.

Also notable is the increase in disagreement between LIME and SHAP at increasing complexities. While LIME and SHAP had an average signed rank agreement of 0.820 at 11 LLMs, this deteriorates to only 0.244 at 634 LLMs. This disagreement is very concerning, especially since there are no well-established methods that researchers consistently employ to resolve these disagreements [7]. The fact that neither LIME nor SHAP consistently outperforms the other across all complexities further exacerbates the issue with disagreement.

Note that this toy case utilizes a small neural network with only 2 layers of 20 nodes. ML practitioners typically use complex black-box neural networks with multiple layers of more than a hundred nodes (and hundreds of LLMs as a result), which can obtain very good performance solely based on accuracy or AUC. However, LIME and SHAP's lack of fidelity can become an even bigger problem that can be very difficult to resolve for these complex models, especially since exact interpretations will not be available for black-box models.

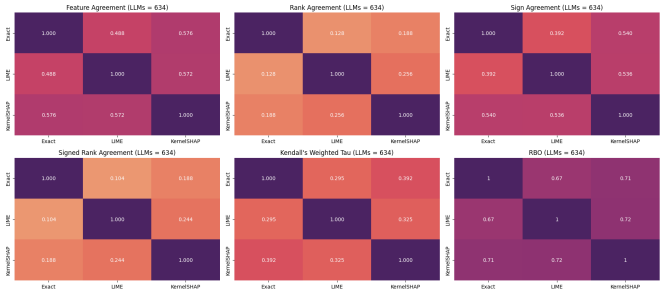


(a) Graph of metrics vs number of LLMs. Metrics tend to monotonically decrease as of LLMs increases.



(b) 11 LLMs

(c) 305 LLMs



(d) 634 LLMs

Figure 2. Average fidelity metrics for exact interpretation, LIME, and SHAP for varying LLM counts of ReLU-DNN models trained on the German Credit Dataset. Notice that all metrics decrease at increasing LLM counts, indicative of decreasing fidelity to exact interpretations and decreasing disagreement between explainability methods.

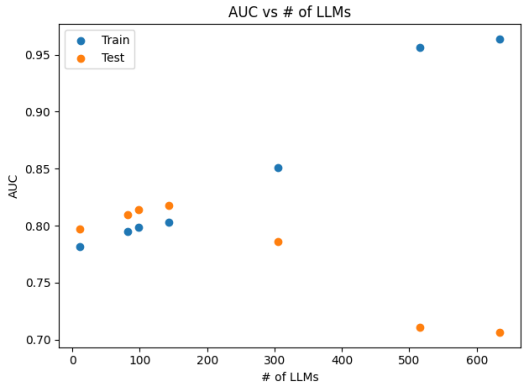


Figure 3. Train and Test AUC of the model after 20 epochs vs. the number of LLMs. As the model becomes more complex, the train AUC increases. After around 150 LLMs, the gap between train and test AUC increases, indicating overfitting. Normally, researchers select a complexity that maximizes test AUC while minimizing the gap between train and test.

5.2. All Models

Datasets			
Dataset	Number of Attributes	Number of Examples	Task
Abalone	8	4177	Regression
Aquatic Toxicity	9	546	Regression
Bike Sharing	16	17389	Regression
California Housing	9	2640	Regression
Steel Industry	11	35040	Regression
Bank Marketing	17	45211	Classification
Breast Cancer	32	569	Classification
Census Income	14	48842	Classification
German Credit	20	1000	Classification
MAGIC Gamma	11	19020	Classification

Table 1. Datasets used in the experiment. Datasets span fields ranging from finance, science, medicine, policy, and more.

We train assorted inherently interpretable models on a range of 5 classification and 5 regression datasets from the UCI Dataset Repository, a repository of real-world datasets [5]. The set of inherently interpretable models we use includes GAM, EBM, and GAMI-Net, and ReLU-DNN (from last section), spanning a wide range of different decision boundaries.

For GAM, we use a spline order of 3, number of knots to 20, smoothness to 0.6, and maximum number of iterations to 100. We set the number of interactions to 10 for EBM and GAMI-Net. For GAMI-Net, we set the size of the subnets for main effects and interactions to 1 and 2 layers of 20 nodes each, respectively. We use the same hyperparameters for the ReLU-DNN as last section, with 2 layers of 20 nodes each and an ℓ_1 regularization of 8e-4.

We train the models until they converge. We harness the inherently interpretable nature of these models to obtain ground truth feature importance rankings for 50 random samples from the test set. Then, we utilize LIME and SHAP to obtain post-hoc feature importance rankings for those same 50 random examples. We compare the explanations to our exact interpretations, allowing for us to gauge the local fidelity of each method.

Our results (Table 3) show that LIME and SHAP have the highest fidelity when used on ReLU-DNN on almost all datasets. The explanation methods' high fidelity on ReLU-DNNs could be due to the piecewise linear nature of the decision boundary of ReLU-DNNs, as compared to a more continuous, smooth one for GAM and GAMI-Net, and a stepwise one for EBM [19] [9] [23].

The fidelity of LIME and SHAP largely depends on the model and dataset used. Notice that the fidelity varies widely for specific models across datasets. For example, LIME and SHAP both do the best job describing GAMI-Net for the Census dataset (Kendall's Weighted Tau of 0.713 and 0.762), yet do comparatively worse for the German Credit dataset (Kendall's Weighted Tau of 0.387 and 0.338). Additionally, neither LIME nor SHAP consistently outperforms the other in fidelity across datasets and models. This is another example of the disagreement problem, showing that it is never fully correct to rely on a single explanation method for a certain scenario [7].

Without access to exact interpretations, there would be no concrete way to gauge whether LIME or SHAP would be faithful to an particular black box model trained on a particular real-world dataset. However, this work demonstrates that less complex, piecewise linear models like small ReLU deep neural networks may be a good choice should utilizing a black-box model be inevitable.

6. Conclusion

In high-risk applications, it is vital for post-hoc explainability methods to grant faithful explanations. An inconsistent or non-exact explanation can mislead researchers and clients, leading to costly and unethical consequences, reducing trust in the explanation and subsequently the model itself [15]. As such, it is important to utilize explanations that one can rely upon to be accurate and consistent.

The inconsistency of LIME and SHAP shown in this paper prove that they should not be relied upon to explain decisions of complex black-box models for high risk applications, where decisions are very important. The performance of LIME and SHAP heavily depends on the dataset and model used. It also depends on model complexity, with more complex models facing a rapid decay in fidelity. The

Weighted Kendall's Tau (LIME)					Weighted Kendall's Tau (KernelSHAP)				
	GAM	EBM	GAMI-Net	ReLU-DNN		GAM	EBM	GAMI-Net	ReLU-DNN
Abalone	0.431	0.398	0.563	**0.827**	Abalone	0.427	0.48	0.673	**0.808**
Bike Sharing	0.266	0.544	0.454	**0.864**	Bike Sharing	0.34	0.729	0.643	**0.865**
Aquatic Toxicity	-0.041	0.281	0.395	**0.998**	Aquatic Toxicity	-0.01	0.747	0.7	**0.963**
California Housing	0.384	0.456	0.621	**0.872**	California Housing	0.385	0.646	0.788	**0.792**
Steel Industry	0.583	0.712	0.709	**0.755**	Steel Industry	0.768	0.871	0.888	**0.913**
Bank Marketing	0.305	0.535	0.709	**0.77**	Bank Marketing	0.48	**0.813**	0.692	0.784
Breast Cancer	N/A	0.501	N/A	**0.785**	Breast Cancer	N/A	0.615	N/A	**0.639**
Census	0.35	0.62	**0.713**	0.56	Census	0.283	0.663	**0.762**	0.535
German Credit	0.344	0.532	0.387	**0.867**	German Credit	0.444	0.789	0.338	**0.865**
MAGIC Gamma	0.273	0.272	0.461	**0.654**	MAGIC Gamma	0.544	0.648	**0.729**	0.689

Table 2. Weighted Kendall's Tau Rank Correlation for LIME and SHAP with Exact Interpretation. Weighted Kendall's Tau ranges from -1 to 1. N/A indicates model did not converge. LIME and SHAP tend to be most faithful when used to interpret a ReLU-DNN. Variances vary from 9.21e-5 to 0.34.

RBO (LIME)					RBO (KernelSHAP)				
	GAM	EBM	GAMI-Net	ReLU-DNN		GAM	EBM	GAMI-Net	ReLU-DNN
Abalone	0.809	0.728	**0.865**	0.827	Abalone	0.801	0.741	**0.899**	0.819
Bike Sharing	0.688	0.809	0.778	**0.959**	Bike Sharing	0.774	0.909	0.872	**0.935**
Aquatic Toxicity	0.598	0.707	0.793	**0.995**	Aquatic Toxicity	0.585	0.904	0.89	**0.972**
California Housing	0.747	0.762	0.857	**0.93**	California Housing	0.75	0.808	0.922	**0.929**
Steel Industry	0.764	0.768	0.879	**0.953**	Steel Industry	0.754	0.824	0.929	**0.966**
Bank Marketing	0.707	0.81	0.759	**0.885**	Bank Marketing	0.749	**0.875**	0.94	0.844
Breast Cancer	N/A	0.746	N/A	**0.933**	Breast Cancer	N/A	0.794	N/A	**0.866**
Census	0.648	0.821	**0.841**	0.813	Census	0.662	0.84	**0.876**	0.807
German Credit	0.678	0.788	0.797	**0.937**	German Credit	0.717	0.899	0.793	**0.936**
Magic Gamma	0.605	0.704	0.776	**0.829**	Magic Gamma	0.699	0.858	**0.899**	0.808

Table 3. RBO Rank Correlation for LIME and SHAP with Exact Interpretation. RBO ranges from 0 to 1. N/A indicates model did not converge. LIME and SHAP tend to be most faithful when used to interpret a ReLU-DNN. Variances vary from 6.58e-4 to 0.0395.

presence of these numerous factors demonstrates that these approximation-based methods should not be trusted to exactly explain predictions after a model is trained.

A more reliable approach to interpreting models is to use glass-box inherently interpretable models. These models can not only obtain a satisfactory accuracy/AUC, but also are able to generate an exact feature importance ranking that can be treated as the ground truth, removing all approximations and guesswork involved. Researchers can utilize these exact interpretations to further validate and improve the model through incorporation of business intuition and additional knowledge.

Furthermore, our work further reinforces the fact that the simpler the model, the higher the interpretability. Therefore, a future direction of work should be to create simple, easy to interpret models that can achieve optimal performance comparable to state-of-the-art black-box models.

References

[1] Chirag Agarwal, Eshika Saxena, Satyapriya Krishna, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations, 2022. 3

[2] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018. 2

[3] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. 2017. 2

[4] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery. 2

[5] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. 4, 6

[6] Sara Hooker, Dumitru Erhan, Pieter jan Kindermans, and Been Kim. Evaluating feature importance estimates. *arXiv*, 2018. 3

[7] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *CoRR*, abs/2202.01602, 2022. 3, 4, 6

[8] Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. *CoRR*, abs/2106.12543, 2021. 3

[9] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 623–631, New York, NY, USA, 2013. Association for Computing Machinery. 6

[10] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. 2

[11] Andreas Messalas, Christos Makris, and Yannis Kanellopoulos. Model-agnostic interpretability with shapley values. 07 2019. 2, 3

[12] Christoph Molnar. *Interpretable machine learning: A guide for making Black Box models interpretable*. Lulu, 2019. 2

[13] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019. 1

[14] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. 2

[15] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. 2018. 1, 2, 3, 6

[16] Cynthia Rudin and Joanna Radin. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, 1(2), nov 22 2019. https://hdsr.mitpress.mit.edu/pub/f9kuryi8. 1

[17] Grace S. Shieh. A weighted kendall's tau statistic. *Statistics Probability Letters*, 39(1):17–24, 1998. 4

[18] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. How can we fool LIME and shap? adversarial attacks on post hoc explanation methods. *CoRR*, abs/1911.02508, 2019. 1

[19] Agus Sudjianto, William Knauth, Rahul Singh, Zebin Yang, and Aijun Zhang. Unwrapping the black box of deep relu networks: Interpretability, diagnostics, and simplification. *CoRR*, abs/2011.04041, 2020. 3, 4, 6

[20] Agus Sudjianto and Aijun Zhang. Designing inherently interpretable machine learning models. *CoRR*, abs/2111.01743, 2021. 3

[21] Agus Sudjianto, Aijun Zhang, Zebin Yang, Yu Su, Ningzhou Zeng, and Nair Vijay. Piml: A python toolbox for interpretable machine learning model development and validation. *To appear*, 2022. 2

[22] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28:20, 11 2010. 4

[23] Zebin Yang, Aijun Zhang, and Agus Sudjianto. Gami-net: An explainable neural network based on generalized additive models with structured interactions, 2020. 1, 6

[24] Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168:022022, 02 2019. 4