The fidelity of global surrogates in interpretable Machine Learning

Carel Schwartzenberg[1], Tom van Engers[1], and Yuan Li[2]

[1] University of Amsterdam, Amsterdam, The Netherlands
carel_schwartz@hotmail.com
vanEngers@uva.nl
[2] ING, Amsterdam, The Netherlands
yuan.li@ing.com

Abstract. In this paper, we focus on an interpretable machine learning technique that has increasingly gained attention as of late, named 'Global Surrogates'. When using a global surrogate, an interpretable 'white-box' model is trained on a less-interpretable, but more accurate, 'black-box' model. The original black-box is used to make the predictions, while the white-box surrogate is used to understand the decision making process. A potential problem with Global Surrogates is the fidelity of the white-box surrogate to the original black-box model. To research the fidelity of global surrogates, we perform three experiments. In the first experiment, we find that the Spearman Correlation is the most appropriate metric to measure the fidelity of surrogates. From the results in the second experiment, we find that Logistic Rule Regression (LRR) and RuleFit, two rule ensembles, consistently show high fidelity. Also, we conclude that the fidelity of the different classes of surrogate models depends quite heavily on the type of original black box. Finally, when we look into the fidelity-interpretability trade-off of global surrogates in the third experiment, we conclude that LRR, RuleFit and decision trees perform well in terms of their fidelity-interpretability trade-off.

Keywords: Interpretable Machine Learning · Global Surrogates · Fidelity.

1 Introduction

Currently, increasingly more decisions are being made with the help of Machine Learning models. A lot of focus is on making these decisions with the highest accuracy possible. However, this focus on high accuracy causes algorithms to become increasingly complex [18]. This increase in complexity comes at a price: models are gradually becoming less and less interpretable [5]. Even though theoretically, the calculations that lead to a certain decision are known, it has gradually become more difficult to explain what the exact cause of a certain prediction is.

Multiple stakeholders benefit from insight into the decisions made by Machine learning models. These stakeholders can be categorized into three main groups

[16]: The people using the model to make decisions, the developers of a model and potentially also the human subjects of a model.

The user of a model can benefit from model explanations by gaining new insights into the task the model is used for. Also, relevant explanations can increase the user's trust in the model, allowing the user to rely more confidently on the model. The creators of a model benefit from model interpretability because explanations are a tool to evaluate a model. For example, model interpretability might allow a developer to find that a model makes illogical decisions or that a model is unintentionally biased. Finally, the human subjects of a model might gain from explanations by learning more about the decision being made about them. Additionally, a model that makes decisions about human subjects requires interpretability by law, as defined in the GDPR.

A wide array of techniques has been developed to tackle the challenge of interpreting complex machine learning models [9]. In the research in this paper, the focus is on one such machine learning interpretability method, called 'Global Surrogates'. In Global Surrogates, an interpretable surrogate "white-box" model is trained on the predictions of an existing less-interpretable "black-box" model, in order to interpret the predictions made by the black-box.

Global surrogates have gained popularity in recent research. Surrogate models have been proposed as "well suited to verify a system and detect failures" by interpretability research [16]. Also, the European banking authority [6] includes surrogates as one of their primary examples of interpretability techniques. Furthermore, research by the data science industry sees surrogates as "appropriate for data scientist entrusted with model development" [2].

For an interpretability method to indeed perform well, two main aspects are important. Firstly, the explanations given by the interpretability method need to be understandable and easy to interpret. For global surrogates this is the case; White-boxes, which are used as the surrogates, are defined by their interpretable nature. Secondly, the explanations offered by interpretability methods need to correctly explain the original black-box model: The explanations need to be of high fidelity to the original black-box model. In contrast to the interpretability of surrogates, the fidelity is a potential weakness.

Interestingly, very little research has been done to show that surrogate models produce high fidelity explanations. Thus, multiple questions arise: Are surrogate models indeed appropriate for regulators and data scientists? How high is the fidelity of explanations given by surrogates really? Can surrogates be relied upon to represent complex black-box models? Is using global surrogates worth it, or should a white-box be used instead?

Based on these questions, we perform a set of three experiments. In the first experiment, we research which metric we should use to measure the fidelity of global surrogates. Then, using this metric, we determine the fidelity of multiple classes of global surrogates on multiple classes of black-boxes in the second experiment. Finally, since the higher fidelity surrogates are expected to be more complex and thus less interpretable, we look into the fidelity-interpretability trade-off of global surrogates in the third experiment.

2 Background and related work

Global surrogates are flexible and widely applicable [12]. They are not only model-agnostic, but also the model that is used as surrogate is interchangeable: any interpretable model can be used as the surrogate.

The idea behind surrogate models is borrowed from engineering: Surrogate modelling is concerned with developing and utilizing cheaper-to-run ”surrogates” of the original simulation model [14]. In other words, if a simulation is computationally very expensive, a cheaper-to-run surrogate is used to approximate the original simulation.

Research on global surrogates for AI interpretability purposes generally has not received a lot of attention yet. Lakkaraju [11] proposes a novel framework to explain a black-box through decision sets, which are trained to be unambiguous, high-fidelity and interpretable. Ribeiro [15] has done research on explaining complex models through the use of Anchors. In this research, Anchors are if-then rules that sufficiently “anchor” the original prediction locally. On a different note, Bastani [3] looks at the construction of decision trees from a black box. Kuttichira [10] also looks into the approximation of a black-box with a decision tree and proposes a novel way of training the decision tree, making sure to stay close to the original model.

3 Selecting the right fidelity metric

Machine learning theory describes a wide array of metrics that can measure the quality of a model. The available metrics can roughly be divided into two subclasses: categorical metrics and continuous metrics. Categorical metrics measure whether instances have a correct categorical classification, while continuous metrics measure the relatedness of the continuous output of a model to a continuous training label. In general, categorical metrics are used for classification purposes, while continuous metrics are used for regression purposes.

In the context of surrogate classification models, both categorical metrics and continuous metrics can be used. We can measure how many of the categorical classification values of the surrogate match the categorical classifications values of the black-box and thus use a categorical metric. However, we can also measure how far the continuous outputs of the models are apart and thus use a continuous metric.

The question arises which group of methods better reflects how well a surrogate resembles the original model. Or, more specifically: Which metric better reflects if the surrogate model will correctly explain the decisions made by the original model.

Since continuous values capture more information than categorical values, we expect the continuous metrics to perform better than the categorical metrics.

There also exist a variety of continuous metrics. A relevant distinction we can make with continuous metrics is between distance-based continuous metrics and correlation-based continuous metrics. Distance-based metrics measure how far predictions and labels are apart in the real-valued prediction space, while correlation-based metrics measure how related the predictions and labels are. In the context of fidelity metrics, correlation-based metrics have the advantage of being scale-invariant, which should be a significant advantage when determining the fidelity of surrogates.

3.1 Experiment 1: Fidelity metrics

We will perform an experiment to validate our hypotheses from the previous section and determine which metric is actually best at determining the fidelity of a surrogate.

Datasets Since a single dataset would not be enough to confidently base conclusions on, multiple datasets will be used to evaluate the metrics on. Artificially generated datasets will be used to have a clear grip on the complexity and to be able to generate an infinite variety of datapoints. Initially, data containing relatively few features and simple decision boundaries is used, after which complexity is gradually increased.

Four types of datasets will be generated, each with their own geometrical decision boundary shape. The four geometrical shapes consist of: vertices, Gaussian blobs, circles and moons. The 2D variants of these dataset shapes can be seen in Figure 1.

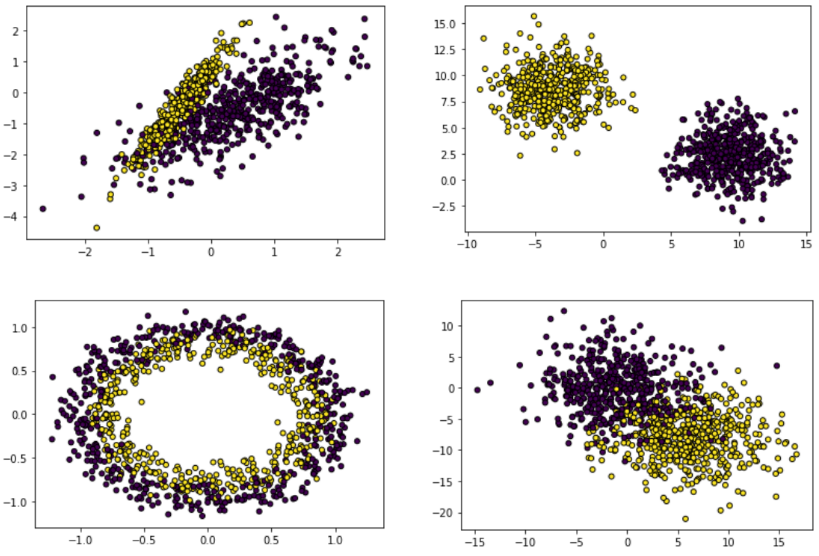

Fig. 1. From left to right, top to bottom: 'Vertices' dataset, 'Blobs' dataset, 'Circles' dataset and 'moons' dataset. The x and y axis represent the feature values, while the color of the points represents the class of the datapoint.

Table 1. The specifications of the generated datasets. The percentage in the last column represent the best average accuracy on the dataset of the models that were trained during this experiment.

Type	# Samples	# Features	# informative	Maximum average classification accuracy
Vertices	1000	2	2	99%
Vertices	1000	2	2	94%
Vertices	1000	10	2	97%
Vertices	1000	10	2	93%
Vertices	1000	10	2	87%
Vertices	1000	10	10	95%
Vertices	1000	10	10	91%
Vertices	1000	20	12	90%
Blobs	1000	2	2	87%
Blobs	1000	10	10	93%
Circles	1000	2	2	92%
Circles	1000	2	2	83%
Circles	1000	2	2	75%
Moons	1000	2	2	90%
Moons	1000	2	2	84%

Multiple key aspects of the generated datasets are changed throughout the experiment, to obtain multiple variations on each dataset type. Aspects that are changed include the number of features and the number of informative features. Also, each of the four types of datasets has specific settings that can be used to tune the class separation and thus to increase or decrease how difficult it is to classify the datapoints. To quantify how challenging each of these datasets is, the 'maximum average classification accuracy' is reported. This is the best average score that the black-box classifiers in the experiment obtained on the dataset.

In the end, a total of 15 dataset configurations is used, as can be seen in Table 1. For each setting of the data generation process, 20 datasets are generated. For each of these datasets, the results are evaluated through 5-fold cross-validation, resulting in a total of 100 training and evaluation cycles per dataset. In the 5-fold cross-validation, the same 4 folds are used for the training of the black-boxes and the white-boxes and the 1 fold is used to apply the fidelity metrics to. A total of 300 datasets is generated (20 per dataset configuration), for each of which we establish which metric agrees on which surrogate is best.

Fidelity measures As fidelity metrics, both categorical and continuous metrics are included, as well as distance- and correlation based-metrics.

Since the focus is on balanced datasets, accuracy suffices as a categorical metric. A second categorical metric that is used, is the Area Under the Curve (AUC). As continuous metrics, both the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) are used. Both these distance-based metrics are included, since it is not clear if squared errors are more impactful than the abso-

Table 2. Overview of the tested fidelity metrics.

Metric	Continuous	Correlation-based
Accuracy	No	No
Area Under the ROC curve (AUC)	Only the predictions	No
Mean Squared Error (MSE)	Yes	No
Mean Absolute Error (MAE)	Yes	No
Spearman Correlation	Yes	Yes
Coefficient of Determination (R^2)	Yes	Yes

lute error in the context of fidelity measurement. The correlation-based metrics of choice are Coefficient of Determination and Spearman correlation. Coefficient of Determination to include a linear correlation measure and Spearman correlation to include a non-linear correlation measure. These six metrics and their characteristics are listed in Table 2.

The accuracy is measured using the output of the original model as labels and the labels produced by the surrogate as predictions. AUC is measured using the binary output of the black box as targets and the continuous output of the surrogate as predictions. The MSE, MAE, Spearman Correlation and Coefficient of Determination are measured using the continuous "probabilistic" output of the original model as labels and the continuous output of the surrogate as predictions. We used these metrics to objectively compare the quality of the surrogate with the original model, and to compare surrogates amongst each others. Which model to choose depends on the specific priorities of the task.

Models Selecting surrogates that are relatively similar in terms of their complexity should lead to more meaningful results. Additionally, surrogate models with both similar and dissimilar inner-working should be included.

For the black-boxes, Random Forests and Neural Networks are selected. The number of estimators for the Random Forest was set to one hundred and Gini was used as the quality criterion. A Neural Network structure with 3 layers was chosen, with a hidden layer width of two times the number of input features. As surrogates, Logistic regression and the rule-ensemble named RuleFit [8] are chosen. The Logistic regression uses SAGA as solver, while RuleFit uses the standard settings as mentioned in the paper by Friedman. The Random Forest and RuleFit are theoretically more related in terms of inner-workings, just like the Neural Network and Logistic regression. Selecting these models leads to a total of four surrogates, two per black box.

Asserting which surrogate is truly best: PMI and ALE To determine which fidelity metric is best at comparing the surrogate to the original model, we need to in some way assert how related the surrogate and original model really are. To do this, we introduce two additional measures that compare the way the surrogate and the original model process the input information into a prediction.

The Permutation Feature Importance (PMI) [7] is a basic and relatively well-known method to measure the importance of each of the features of a machine learning model. PMI will be used to determine how similar the input feature importance's of the surrogate and the original black-box are. The PMI is computed for each of the features of the black box and the surrogate, after which the mean absolute difference is calculated. The result is a single measure that indicates how similar the feature importance's of the surrogate and the original model are.

The Accumulated Local Effect (ALE) [1] plots describe how features influence the predictions made by a machine learning model. The ALE plots of the surrogate and the original black-box will be used to compare the marginal input features effects of the surrogates to those of the original black-boxes. The similarity between the ALE plots of the surrogate and the ALE plots of the black-box is compared through the Mean Squared Error between the ALE measures of both models, resulting in a single measure that indicates the similarity of the marginal feature effects.

The PMI and ALE measures are relatively expensive to calculate and these measures are not as easy to apply to every machine learning model out there. For this reason, the PMI and ALE measures can not simply be used as fidelity measures, but have to be used as guidance to point out which metric should in fact be used as fidelity measure.

3.2 Results of experiment 1

Since there are 15 dataset configurations for which results are gathered, results are too numerous to present for each dataset separately. To solve this, we combined the results into a legible format as follows: For each black-box, two surrogates compete to be the best performing surrogate. Depending on which metric we look at, a different surrogate might seem to perform better. For each of the black-boxes, we determine which metrics agree on which surrogate is best and which do not. The total percentage of times these metrics agree with each other are shown in Table 3. The ALE and PMI metrics give an indication of which surrogate really performs best, i.e. which surrogate is actually closest to the original black-box. Thus, the metrics that frequently “agree” with the ALE and PMI measures can be seen as well-performing measures.

3.3 Conclusion on the fidelity metric

The results show that the absolute difference in relatedness fraction is not very big for most metrics. It should be clarified however that this is partly due to the fact that in some cases, the difference in performance between the two surrogates is significant, while in other cases the difference in performance is small. This means that it is either very clear which surrogate is best and most (if not all) metrics will agree on which surrogate is best, or the difference in performance is very small, resulting in metrics that do not agree on which surrogate

Table 3. The fraction of times the metrics and PMI/ALE agree on which surrogate performs best.

	Acc.	MSE	MAE	R^2	AUC	SpCor
Accuracy (Acc.)	1.00	0.677	0.664	0.677	0.761	0.654
Mean Squared Er. (MSE)	0.677	1.00	0.941	1.000	0.759	0.827
Mean Absolute Error (MAE)	0.664	0.941	1.00	0.941	0.743	0.796
Coefficient of Determination (R^2)	0.677	1.000	0.941	1.00	0.759	0.827
Area Under the Curve (AUC)	0.761	0.759	0.743	0.759	1.00	0.764
Spearman Correlation (SpCor)	0.654	0.827	0.796	0.827	0.764	1.00
Permutation Importance (PMI)	0.648	0.675	0.666	0.675	0.713	**0.720**
Accumulated Locale Effects (ALE)	0.659	0.711	0.723	0.711	0.738	**0.788**

is best. On top of this, for this experiment, "classifying" which surrogate performs best could be seen as a binary classification task: Per black-box, there are only 2 options. This means that a randomly classifying a surrogate as best produces a baseline of 0.5. Combined with the obvious cases, this leads to an easily achievable relatedness score of about 0.6.

As for the best fidelity metric, Spearman Correlation performs best in terms of PMI- and ALE-relatedness. Based on the theoretical considerations at the start of this chapter and the results of the experiment, we thus conclude that Spearman Correlation is the most appropriate metric to use, or at least out of the metrics that were evaluated.

4 Experimental setup

Now that we have found an appropriate fidelity metric, we define a second and third experiment. In the second experiment, we use the Spearman Correlation to measure the fidelity of surrogates. In the third experiment, we look at the fidelity-interpretability trade-off of the surrogates.

4.1 Experiment 2: The fidelity of global surrogates

In the second experiment, we determine how high the fidelity of global surrogates is in general, as well as which classes of surrogates align especially well with certain classes of black-boxes.

Datasets and models To obtain reliable results, we again use a variety of datasets. This time, datasets with real data are used in addition to the generated datasets that were used in experiment 1. This is done to ensure the results are applicable to real-life data and applications.

A range of financial and non-financial datasets is selected for the real datasets, with varying amounts of instances and features, as can be seen in Table 4.

The Mushroom and HELOC datasets are balanced, while the Creditcard, Census Income and Statlog datasets are not. These datasets are balanced artificially through under-sampling of the more available class.

Table 4. The datasets that were used, each with their number of instances, number of features and if the datasets is balanced of itself.

Dataset name	Instances	Features	Balanced
UCI mushroom classification dataset	8124	22	Yes
Home Equity Line of Credit (HELOC)	9871	23	Yes
UCI Creditcard Clients	30000	24	No
UCI Census Income	48842	14	No
UCI Statlog	58000	20	No

Again, 5-fold cross-validation is applied to the datasets to ensure stable results. Just like in the first experiment, the four folds are used as training data for both the black-boxes and white boxes, while the last fold is used to measure the fidelity of the surrogate. The training process will be run 10 times for each dataset.

Multiple black-box and white-box machine learning algorithms are selected to be trained on the datasets. Since there exists a near-infinite number of machine learning models, we limit ourselves to black-box models that are popular in industry. For the white-boxes, well-known models are favoured as well, however, we also include white-boxes that are slightly less well-known, but have shown promising predictive performance in literature.

An overview of the models used can be seen in Table 5. As black boxes, Random Forest, AdaBoost, XGBoost, Neural Networks and SVM are selected. These models are widely used in industry, while this selection also includes a variety of inner-workings. The Neural Network contained 4 layers, with a hidden-layer width of two times the number of features.

As the white-boxes, decision trees, logistics regression and Naïve Bayes are generally well-known and widely used. The rule-ensembles RuleFit and Logistic Rule Regression [17](LRR) are included because previous research suggests these models to have a favourable fidelity-interpretability trade-off. Logistic regression again uses SAGA as solver, the decision trees automatically grid-searches for the optimal tree-depth and for Naïve Bayes the standard settings are used. RuleFit and LRR use the standard settings as suggested in their respective papers.

Table 5. The black-box and white-box surrogates that were tested.

Black-boxes models	White-box surrogates
Random Forest	Decision Tree
AdaBoost Classifier	Logistic Regression
GradientBoost classifier	Naïve Bayes
Support Vector Machine	RuleFit
Neural Network	Logistic Rule Regression

4.2 Experiment 3: Fidelity-interpretability trade-off

Selecting a surrogate often entails more than purely the fidelity of that surrogate: The fidelity-interpretability trade-off of the surrogate can also be important. This is why in addition to evaluating the surrogates purely based on their fidelity, the interpretability of the surrogates will also be quantified. This should give a more complete picture with respect to which surrogate is most fit for a certain black-box and task.

To quantify the interpretability of the surrogates, a methodology from a paper by Molnar is used [13]. Molnar quantifies the interpretability of a machine learning model as a combination of three factors: the 'Main Effect Complexity', the 'Interaction Strengths' and the 'Number of Features'. The 'Main Effect Complexity' represents the complexity of the relationships between the input feature and the prediction, the 'Interactions Strengths' represents the strength of the interactions between the features and the 'Number of Features' represents the number of features that have an influence on the outcomes of the model. By re-scaling and combining these three features, Molnar comes to a single interpretability score for each surrogate, where a score of zero means low relative interpretability and a score of three means high relative interpretability.

The datasets and models used in this experiment are the same real datasets that were used in the second experiment, for the obvious reason that this experiment is meant to give insight into the interpretability of the models that were used in the previous experiment. The generated datasets are not expected to offer additional insights into the interpretability of the models, because of the relative simplicity of these datasets.

5 Results

5.1 Experiment 2: Fidelity of surrogates

For each of the five black-boxes, fidelity results of the five surrogates are gathered over all the datasets. In Figure 2, we report how on how many datasets each surrogate performs best per black-box. Notably, the best performing surrogate varies significantly per black-box, but also depending on the dataset. For Random Forests and XGB, RuleFit is the most frequent best performer. For AdaBoost and SVM, Logistic Rule Regression is the most frequent best performer. For Neural Networks, Logistic Regression performs best.

In Figure 3, the average Spearman correlation over all twenty datasets is reported. The average Spearman Correlations tell a similar story as the results in Figure 2: RuleFit and Logistic Rule Regression perform best in general. This time however, Logistic rule regression performs better than Logistic regression on Neural networks. This is because while Logistic regression performs best on most datasets, it performs poorly on the remaining datasets. A main reason for this is that logistic regression is not able to take on circular decision boundaries, causing it to perform badly on the generated circular datasets as well as on the real datasets that require circle-shaped decision boundaries.

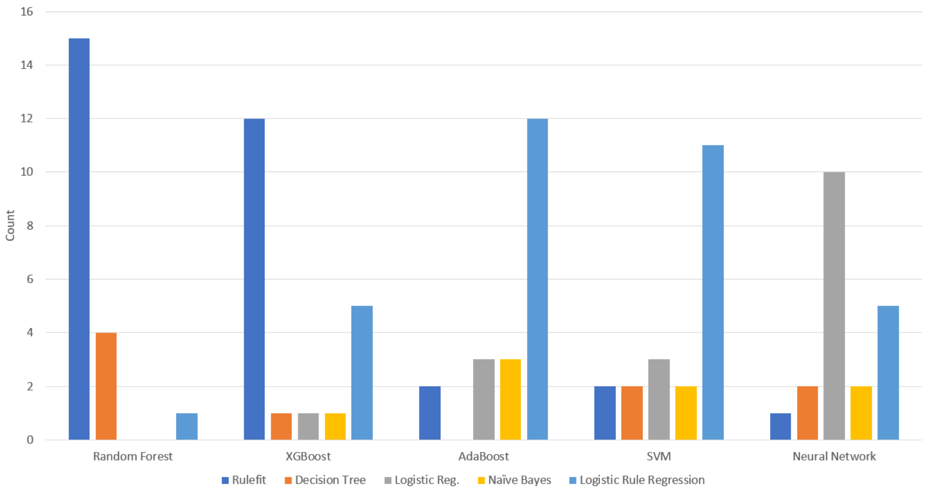

Fig. 2. The frequency a surrogate performs best fidelity-wise on a certain black-box, over all twenty datasets.

5.2 Experiment 3: Interpretability of surrogates

The interpretability experiment has been performed on the exact same models and datasets as the results in the fidelity experiment.

The results can be seen in Figure 4. The interpretability of the surrogates is relatively consistent over the Black-boxes. Overall, decision trees show a high

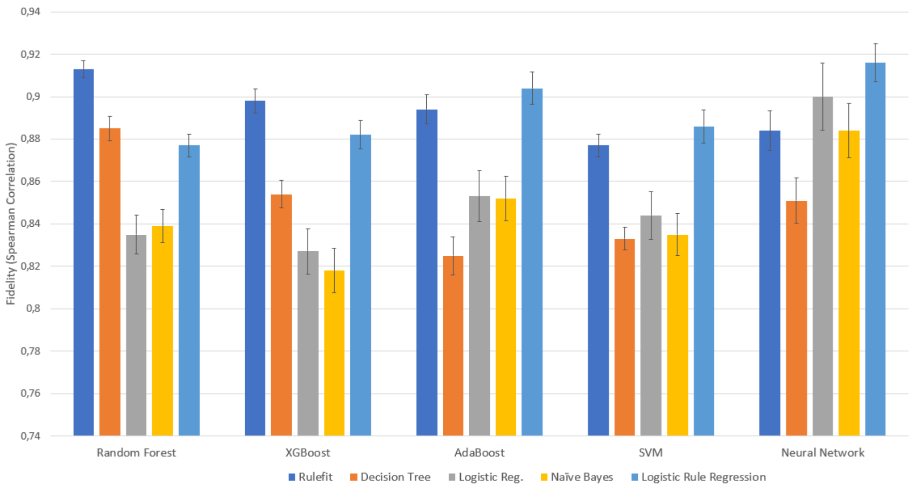

Fig. 3. The fidelity of the five types of surrogates on each of the five types of black-boxes, averaged over all twenty datasets.

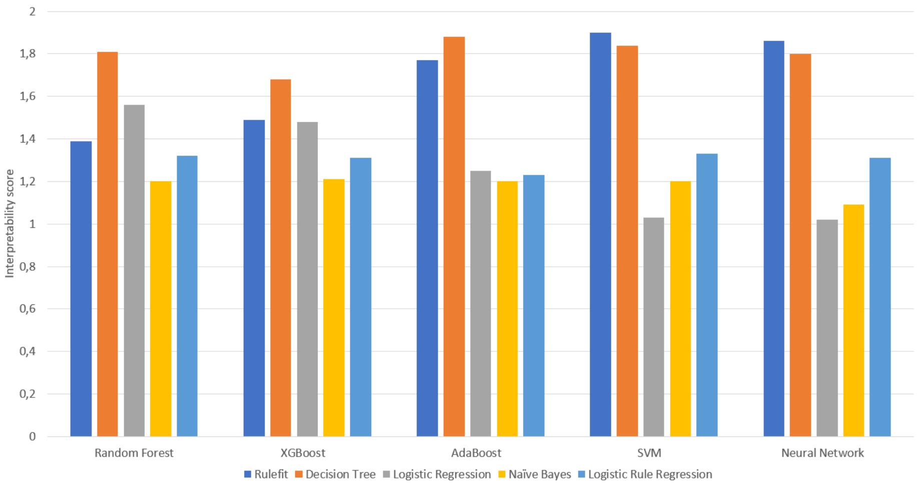

Fig. 4. The interpretability of the five types of surrogates on each of the five types of black-boxes, averaged over the five real datasets. The error bars show the 95% confidence interval.

interpretability score for most black-boxes. RuleFit also performs well, especially on the SVM and Neural Network black-boxes.

If we then combine the fidelity results from experiment 2 and the interpretability results from experiment 3, in both cases using the results of the five real datasets, we produce the fidelity-interpretability plots as shown in Figure 5. For the black-boxes XGBoost, AdaBoost, SVM and Neural Networks, the white-boxes RuleFit and LRR are strong contenders on the Pareto-optimality curve. Decision trees also perform relatively well, especially when focusing on interpretability. For Random Forests, Decision trees trump the other surrogates. It must be noticed that we did not include confidence intervals for the measured interpretability scores. Reason for this is that determining the confidence intervals for Molnar's interpretability score is complicated and exceeds the scope of this paper.

6 Conclusion and future work

In this paper, we investigate how well suited global surrogates are as an AI interpretability method. We performed three experiments: First, we determined which metric is most suitable to measure the fidelity of surrogates. Subsequently, we performed an experiment to determine the fidelity of the surrogates. Finally, an experiment was done to also determine the interpretability of the surrogates.

A variety of white-box surrogate models were trained on a variety of black-boxes, using a multitude of datasets. Based on the theoretical findings and the first experiment, Spearman Correlation appears to be the most appropriate fidelity metric. When we then use the Spearman correlation to measure the fidelity

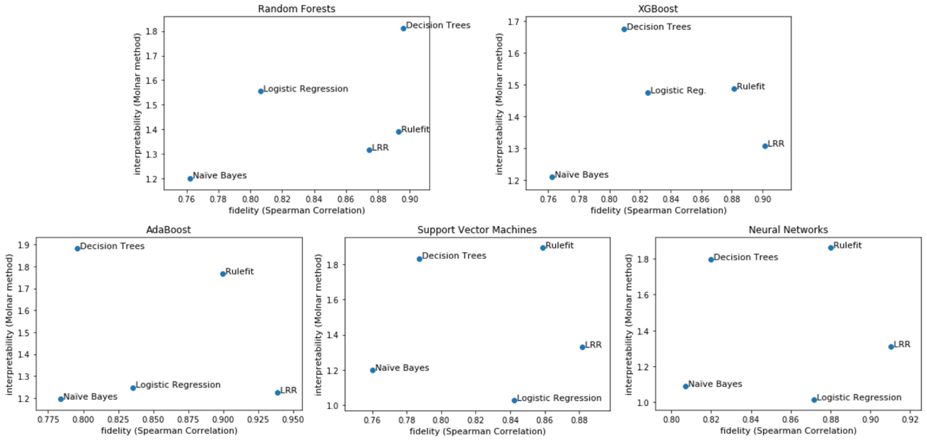

Fig. 5. Fidelity-interpretability trade-off of the surrogates for the five black-boxes. Both the fidelity and interpretability results were averaged over the five real datasets.

of white-box surrogates in experiment 2, we find that the rule-ensembles named Logistic Rule Regression and RuleFit perform well fidelity-wise. The results also show that certain classes of surrogates are better suited for certain classes of black-boxes. The results of the third experiment show that the interpretability of the surrogates is relatively consistent over most black-boxes. In general, Decision trees and RuleFit show high levels of interpretability. If we then plot the fidelity-interpretability trade-off based on the results of experiment 2 and 3, RuleFit, Logistic Rule Regression and Decision trees perform well. Logistic Rule regression does especially well fidelity-wise, while decision trees perform better interpretability-wise.

Back in the introduction, we asked ourselves: “Are surrogate models appropriate for regulators and data scientists?” and “Can surrogates be relied upon to represent complex black-box models?”. We know which surrogates perform best fidelity wise, however, is the level of fidelity of theses surrogates sufficient?

In practice, this is a difficult question to answer. We would say there is not necessarily an a-priori specifiable level of fidelity for a surrogate to be considered reliable, since the required level of fidelity heavily depends on the context in which the interpretability method is applied. For example in healthcare-related applications, surrogate fidelity might be of much higher importance than in more business-related settings, like product recommendations.

We also asked at the start: ”Is using global surrogates worth it, or should a white-box be used instead?”. Most of the time, this comes down to the white-box vs black-box discussion. While black-boxes generally outperform white-boxes in terms of predictive performance, this heavily depends on the dataset and the specific white-box and black-box. Obviously, in cases where a certain white-box performs as well as the best performing black-box, using the white-box is an easy choice. Especially LRR and RuleFit will in some cases perform just as well as the black-boxes. In most cases, however, the better black-boxes will outperform the white-boxes. In general, this out-performance will entail a few percentage points of accuracy. Selecting the right model for the task here, white-box or black-box, should again be context-dependent: Is a slight increase in accuracy more valuable, or directly available and accurate explanations?

We would thus emphasise that the context in which interpretability is needed, is key. On most datasets and black-boxes, the best performing surrogates reach Spearman Correlation scores of at least 0.9, which translate to AUC scores of at least 0.99. In general, this should lead to fairly correct, high-fidelity explanations.

6.1 Future work

Firstly, the research in this paper specifically focuses on balanced binary classification datasets. However, many real-life datasets do not fit this specification. Therefore, future work includes fidelity research on a wider variety of datasets. Many of the methodologies used in this paper still apply to non-binary, unbalanced and/or regression datasets, however, the related results and conclusions might be different and are therefore worth looking into.

Secondly, the selection of datasets in this paper is focused on data with a relatively low number of features. While many real-life datasets will also contain relatively few features (two dozen at a maximum), black-boxes (and especially deep Neural Networks) are especially well-suited for data with a higher numbers of features. A focus of future work could be to investigate if the conclusions of the research in this paper hold for higher numbers of features. This would also give insight into if global surrogate techniques can be applied to Computer Vision or Natural Language Processing use-cases.

Thirdly, the standard methodology to train a global surrogate is used in this paper: The surrogates are trained directly on the outcomes of the original black-box. Alternative surrogate training methods might however certainly yield better results. Therefore, a second direction of future work could be to investigate ways to increase the fidelity of global surrogates. One such alternative method would be the training strategy used in the ProfWeight algorithm [4]. In ProfWeight, the importance of each training sample is weighted by the performance of the original black-box on that sample, instead of weighting each sample in the training set equally. The ProfWeight paper reports a significant increase in surrogate fidelity using this method.

Fourthly, in the application of global surrogates, it is generally assumed that the same surrogate is used to explain every decision made by the black-box. However, it could be that a certain surrogate shows higher fidelity on some subsections of the data, while it shows lower fidelity on other subsections of the data. Therefore, another potentially interesting direction of future research would be to look into the performance of surrogates on the multitude of subsections of datasets.

Fifthly and finally, since the interpretability of a machine learning model has no clear (mathematical) definition, the literature on the subject has a hard time defining robust interpretability quantification methods. Molnar's interpretability quantification method, which was used in this paper, is one of the more robust options. The methodology however does have its limitations: It focuses on the functional complexity of a model, instead of the degree of interpretability of a model to a human. Also, Molnar's method doesn't contain a clear way to determine confidence intervals for the estimated level of interpretability of the models. Future work could look into different, potentially more human-focused interpretability quantification methods. Specifically, more focus could be put on the level of interpretability of a variety of knowledge representation formats that can be used to represent the surrogate's decisions. This should lead to a different perspective on the fidelity-interpretability trade-off of global surrogates.

References

1. Apley, D.W.: Visualizing the effects of predictor variables in black box supervised learning models. arXiv: Methodology (2016)
2. Arya, V., Bellamy, R.K.E., Chen, P.Y.: One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. ArXiv **1909.03012** (2019)
3. Bastani, O., Kim, C., Bastani, H.: Interpreting blackbox models via model extraction. ArXiv **1705.08504** (2017)
4. Dhurandhar, A., Shanmugam, K., Luss, R., Olsen, P.A.: Improving simple models with confidence profiles. ArXiv **1807.07506** (2018)
5. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Communications of the ACM **63**(1), 68–77 (2019). https://doi.org/10.1145/3359786
6. EBA: European banking authority report on big data and advanced analytics (2020)
7. Fisher, A.J., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. **20**, 177:1–177:81 (2019)
8. Friedman, J.H., Popescu, B.E.: Predictive learning via rule ensembles (2008)
9. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An approach to evaluating interpretability of machine learning. CoRR **abs/1806.00069** (2018), http://arxiv.org/abs/1806.00069
10. Kuttichira, D.P., Gupta, S.K., Li, C., Rana, S., Venkatesh, S.: Explaining black-box models using interpretable surrogates. In: PRICAI (2019)
11. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable explorable approximations of black box models. ArXiv **1707.01154** (2017)
12. Molnar, C.: Interpretable Machine Learning, A Guide for Making Black Box Models Explainable (2019), https://christophm.github.io/interpretable-ml-book/
13. Molnar, C., Casalicchio, G.: Quantifying interpretability of arbitrary machine learning models through functional decomposition. ArXiv **1904.03867** (2019)
14. Razavi, S., Tolson, B.A., Burn, D.H.: Review of surrogate modeling in water resources. Water Resources Research **48** (2012)
15. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: AAAI (2018)
16. Ribera, M., Lapedriza, A.: Can we do better explanations? a proposal of user-centered explainable ai. In: IUI Workshops (2019)
17. Wei, D., Dash, S., Gao, T., Günlük, O.: Generalized linear rule models. ArXiv **1906.01761** (2019)
18. Yao, Y., Xiao, Z., Wang, B., Viswanath, B., Zheng, H., Zhao, B.Y.: Complexity vs. performance. Proceedings of the 2017 Internet Measurement Conference (2017)