Data Mining and Knowledge Discovery (2024) 38:237–280
https://doi.org/10.1007/s10618-023-00971-3

Can local explanation techniques explain linear additive models?

Amir Hossein Akhavan Rahnama[1] · Judith Bütepage[1] · Pierre Geurts[2] · Henrik Boström[1]

Received: 28 November 2022 / Accepted: 27 July 2023 / Published online: 19 September 2023

Abstract
Local model-agnostic additive explanation techniques decompose the predicted output of a black-box model into additive feature importance scores. Questions have been raised about the accuracy of the produced local additive explanations. We investigate this by studying whether some of the most popular explanation techniques can accurately explain the decisions of linear additive models. We show that even though the explanations generated by these techniques are linear additives, they can fail to provide accurate explanations when explaining linear additive models. In the experiments, we measure the accuracy of additive explanations, as produced by, e.g., LIME and SHAP, along with the non-additive explanations of Local Permutation Importance (LPI) when explaining Linear and Logistic Regression and Gaussian naive Bayes models over 40 tabular datasets. We also investigate the degree to which different factors, such as the number of numerical or categorical or correlated features, the predictive performance of the black-box model, explanation sample size, similarity metric, and the pre-processing technique used on the dataset can directly affect the accuracy of local explanations.

Keywords Explainable machine learning · Local model-agnostic explanations · LIME · SHAP

1 Introduction

As machine learning models have become more complex, the need for techniques that explain the decision-making process of the *black-box models* has grown (Molnar et al. 2022; Rudin 2018; Ribeiro et al. 2016). To make the decision-making process more accessible to humans, explanation techniques can be used to estimate the importance of features of the data to the model's predicted output. Explanation

Responsible editor: Charalampos Tsourakakis.

Extended author information available on the last page of the article techniques extract the information from the black-box model in a post-hoc manner, i.e., based on the model that is already trained on a dataset (Molnar et al. 2022; Montavon et al. 2018). Explanations can have different representations, such as logic rules (Ribeiro et al. 2018), example-based explanations (van der Waa et al. 2021) and, arguably the most popular type of explanation in the literature, feature attributions (Ribeiro et al. 2016; Lundberg and Lee 2017). The focus of our study is feature attribution techniques that can explain the predicted output of any class of machine learning models for a single instance in a dataset. These techniques can be further divided into additive vs. non-additive explanations. The sum of importance scores in a *local additive* explanation equals the predicted output score for the explained instance (Lundberg and Lee 2017).

In Rudin (2018), the author argues that local explanations,[1] such as LIME and SHAP, can be inaccurate and should not be used in high-stake decision-making domains. The main underlying reason for this argument is the infidelity (inaccuracy) of explanations. The study includes examples of the failure cases of explanations in object detection scenarios. Similarly, other studies have evaluated local explanations of neural networks trained on text and image data[2]. However, the majority of the datasets in high-stake decision-making scenarios, e.g., health and diagnostic (Hakkoum et al. 2022), law (Wang et al. 2022) and so forth are tabular datasets. The question of the explanation accuracy of models used in these high-stake domains is of critical importance.

In this work, we propose to evaluate explanation techniques not when explaining black-box models but when explaining linear additive models, such as Linear and Logistic regression trained on tabular datasets. In particular, we investigate whether local model-agnostic additive explanations can explain linear additive models with high explanation accuracy. We demonstrate how to extract Model-Intrinsic Additive Scores (MIAS) from these models that can directly be compared to the feature importance scores generated with a local explanation technique (see Sect. 4.1 for the definition of explanation accuracy that we employ in this study and more details).

One might wonder whether the answer to this research question is essential since linear additive models are intrinsically interpretable and are not representative black boxes. We show that since we can extract *local* ground truth importance scores from linear additive models and measure the explanation accuracy directly, testing the ability to explain these models can serve as a sanity check for evaluating local additive explanation techniques. This evaluation should be the first step when designing new evaluation techniques. If an explanation technique cannot accurately explain a simple model, we cannot trust its explanations of black-box models either.

One of the most important aspects of evaluating local explanations is understanding the factors affecting the explanation's accuracy. Some studies have studied factors that can cause the accuracy of local explanations to decrease (Molnar et al. 2022; Gosiewska and Biecek 2019). The authors have pointed out three

[1] For brevity, we sometimes refer to local additive model-agnostic explanations as local explanations or just explanations in this study.

[2] See Sect. 2 for details.

main factors that can affect the accuracy of local model-agnostic explanations: (1) The presence of categorical features, (2) The presence of correlated features in the dataset, and (3) Explaining models with low predictive performance. Even though these limitations are frequently mentioned in the literature (Molnar et al. 2022; Gosiewska and Biecek 2019; Guidotti 2021) , the investigations are not conclusive for tabular datasets, and the degree to which these factors can contribute to the accuracy of explanations is not well studied beyond simple cases of synthetic datasets. When explaining linear additive models, our study investigates the aforementioned factors' effect on synthetic and real tabular datasets. Moreover, we show that the accuracy of local explanations is affected by other factors, e.g., the explanation sample size, the choice of similarity metric, and the preposessing technique used on the dataset.

In our investigation, two widely used techniques for generating local additive model-agnostic explanations, Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016) and SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017), are evaluated along with the non-additive explanation technique Local Permutation Importance (LPI) (Casalicchio et al. 2018). The reason for including LPI in our study is to examine how a technique that does not rely on the "additivity" of the local explanations can still produce accurate explanations for linear additive models. We evaluate the explanation accuracy for regression and classification tasks, using linear regression models for the former and logistic regression and Gaussian Naive Bayes models for the latter.

In conclusion, our contributions are

1. We present a novel principled method to extract the local ground truth model-intrinsic importance scores from additive terms in linear additive models.
2. Based on these scores, we describe how to measure the explanation accuracy of local explanation techniques directly, thus providing a sanity check for these methods.
3. Using our proposed accuracy measure, we show that the previously mentioned factors can indeed influence explanation accuracy.

The key findings from the empirical investigations are: (1) LIME and SHAP pass the proposed sanity check for Linear Regression models, (2) The explanation techniques frequently fail the proposed sanity check when explaining Logistic Regression and naive Bayes models, (3) The explanation accuracy of additive explanations of LIME and SHAP is overall larger than for the non-additive local explanations of LPI when explaining linear additive models, (4) In some datasets, LPI explanations are more accurate than explanations of LIME and SHAP when explaining linear additive classification models, even though LPI explanations are not additive (5) All of the aforementioned factors may significantly affect explanation accuracy, however, their effect is largely dependent on the type of model explained and the explanation technique itself, and (6) The most accurate local explanations are not necessarily the the most robust[3] and vice versa.

The rest of the paper is organized as follows. We provide an extensive background on evaluating local explanations in Sect. 2.2. In Sect. 3, we provide a motivating example that shows the limitations of current evaluation measures for evaluating local additive models and highlights the key differences between the proposed evaluation method with other previously proposed approaches. We formally introduce the evaluation method in Sect. 4. In Sect. 5, we empirically study the accuracy of explanation techniques on 40 tabular datasets using the proposed evaluation framework. We discuss the most important findings and the limitations of our study in Sect. 6, and finally, we summarize the main conclusions and point out directions for future research in Sect. 7.

2 Background

Explanation techniques can be divided into global vs. local techniques and model-agnostic vs. model-based techniques. *Global* explanation techniques (Breiman 2001) provide importance scores for features with respect to a dataset (Freitas 2014). *Local* explanation techniques (Ribeiro et al. 2016; Lundberg and Lee 2017) provide importance scores for a prediction of a single instance (Ribeiro et al. 2016). *Model-agnostic* explanation techniques (Ribeiro et al. 2016) can produce explanations for any type of black-box model (Ribeiro et al. 2016). On the other hand, *model-based* explanation techniques (Zeiler and Fergus 2014) are tailored for one type of machine learning model (Montavon et al. 2018). We focus on local model-agnostic explanation techniques. These can be further divided into additive vs. non-additive explanations. The sum of importance scores in a *local additive* explanation equals the predicted output for the explained instance (Lundberg and Lee 2017). In contrast, *local non-additive* explanations do not satisfy the additivity criterion (Lundberg and Lee 2017). Some of the most popular explanation techniques, such as LIME and SHAP, fall into the former category.

In this section, we first formalize local explanations, as produced by LIME, SHAP, and LPI, and then discuss methods to evaluate such explanations.

2.1 Local explanations

We first present a formalization of local additive explanations in our study, i.e. LIME and SHAP, based on the notation used in Lundberg and Lee (2017). As discussed in Sect. 1, in local explanations, a black-box model's predicted output is decomposed into an additive sum of feature importance scores. In simpler words, each feature importance score is the contribution of that feature to the predicted output of the explained model. The formal representation of local explanations, as produced by

[3] See Sect. 2.2.1 for the definition of explanations robustness.

LIME and SHAP, is shown in Eq. 1. In this equation, the black box model f predicted probability for a designated class given instance x is decomposed into an additive sum and ϕ_j is the local feature contribution of feature j and x_j[4] is the value of feature j in x.

$$f(x) = \sum_{j=1}^{M} \phi_j x_j \tag{1}$$

Local Permutation Importance (LPI) is a local non-additive model-agnostic explanation technique. The core idea behind LPI (Casalicchio et al. 2018) is that the importance of a feature can be estimated by the average change of a black-box's predicted output when the value of this feature is replaced by another value. To change the feature value, LPI randomly permutates feature values of a single dimension across all data points in a given dataset.

More formally, LPI is calculated as follows. Let π be a random permutation of the index sequence $\langle 1, \dots, N \rangle$, and let π_i denote the position of index i in π. The importance of feature j at x_n is then defined as:

$$\Phi_n^j = \frac{1}{N} \sum_{k=1}^{N} \left(f(\hat{x}_k) - f(x_n) \right) \tag{2}$$

where $\hat{x}_k$ is defined as follows:

$$\hat{x}_k^l = \begin{cases} x_n^l & l \neq j \\ x_{\pi_k}^j & l = j, \end{cases} \tag{3}$$

where $k \in [1, N]$ and $l \in [1, M]$. In simpler terms, $\hat{x}_k$ is equal to x_n except that the value of the jth feature is replaced by $x_{\pi_k}^j$. It is noteworthy that in our study, $f(x)$ is the log odds ratio prediction function of class c instead of the predicted values $f(x_n)$ for Logistic Regression and Naive Bayes models.

2.2 Evaluating local explanations

The evaluation methods for local explanations are categorized into the human evaluation and functionally grounded evaluation methods (Doshi-Velez and Kim 2017). In human evaluation methods, the accuracy of a local explanation is measured by how accurately human subjects can guess the prediction of black-box models when they have only access to the explanation (Poursabzi-Sangdeh et al. 2021). Since human studies are costly and time-consuming, functionally grounded evaluation methods use different proxies to measure the quality of local explanations. This study focuses on evaluating local explanations using the latter techniques.

[4] In some explanation techniques, such as LIME and SHAP, x_j is replaced by x'_j which is a binary (interpretable) representation of x_j.

Evaluating local explanations using the functionally-grounded method is challenging. We should remember that we need local explanations, or explanations in general, because we do not understand black boxes. For a direct evaluation of explanations, ground truth importance scores are information only directly accessible when we can understand the model. On the other hand, black-box models are models we cannot understand. Because of this, all evaluation methods of local explanations either measure the explanations indirectly, e.g., robustness measures, or they induce further assumption of the data generation process or the model type explained. Because of this, we need to consider that these measures study different characteristics of an explanation.

This section provides a background on each of these evaluation procedures. The majority of studies that have evaluated local explanations have focused on three categories of evaluation procedures: evaluating explanations using robustness measures (Sect. 2.2.1) , using ground truth feature importance scores from synthetic datasets (Sect. 2.2.2) , and using interpretable models (Sect. 2.2.3). Our proposed method belongs to the latter category.

2.2.1 Robustness measures

Most studies on evaluating local explanations, especially for neural network models, use the robustness measures (Fong and Vedaldi 2017; Montavon et al. 2018; Alvarez-Melis and Jaakkola 2018). Robustness measures do not rely on the ground truth importance scores to evaluate explanations (Alvarez-Melis and Jaakkola 2018; Montavon et al. 2018; Adebayo et al. 2018; Lakkaraju et al. 2020; Agarwal et al. 2022). Instead, the main assumption of these measures is that nullifying important (unimportant) features from a local explanation needs to cause large (small) changes in the predicted scores of the black-box models of that instance. In these measures, the black-box model is used as an oracle to extract new prediction scores on the new variation of the explained instance after subsets of its features are nullified. Measures such as faithfulness (Alvarez Melis and Jaakkola 2018), fidelity (Amparore et al. 2021), Prediction Gap on Important Features (PGI), and Prediction Gap on Unimportant Features (PGU) (Agarwal et al. 2022) are all variations of robustness measures. The main reasons behind the popularity of robustness measures are: (1) There is no need to access ground truth importance scores for evaluating local explanations (2) They can evaluate local explanations of arbitrary datasets and explained models..

Our study uses the prevalent Deletion and Preservation robustness measures initially proposed in Fong and Vedaldi (2017); Samek et al. (2016). Our definitions follow the notation from Hsieh et al. (2020). Let $S_r \subset U$ be the set of top-K features ranked in descending order by their absolute importance scores obtained from an explanation technique (K is a hyper-parameter). Let $\bar{S}_r = U \setminus S_r$ where U is the set of all features. Deletion measures the absolute change in a black box's predicted output after replacing feature values in S_r with a baseline value. Similarly, Preservation reflects the absolute change in the predicted output of a black-box model following the replacement of feature values in $\bar{S}_r$ with a baseline. The baseline value can be a binary value or the average value of the corresponding feature in the training or validation set (Fong and Vedaldi 2017). There are no agreed optimal values for these robustness measures. However, a robust explanation should have relatively large Deletion and low Preservation values (Montavon et al. 2018).

Robustness measures are intrinsically prone to have the following limitations: First, since robustness measures are not calculated based on the ground truth importance scores, we cannot argue that robust explanations are *directly* accurate (We show an example of this limitation in Sect. 3). Second, the prediction of an instance after removing its features can cause out-of-distribution predictions or at worst, can turn the instance into an adversarial example. Hence the predictions of the oracle can no longer be trusted to evaluate local explanation. Rahnama and Boström (2019); Hooker et al. (2019); Hsieh et al. (2020). Third, there is a lack of agreement on a unified approach to nullify features (Sturmfels et al. 2020). Fourth, there are no agreements on the most optimal threshold of the magnitude of the change in the predicted probability of the model after (important) unimportant features are removed (Alvarez-Melis and Jaakkola 2018; Sturmfels et al. 2020).

In Hooker et al. (2019), the authors propose an extra step of retraining the model after nullifying important features to avoid the problem of out-of-distribution prediction of the explained model. This is mainly to tackle the second limitation of the robustness measure. In their study, the model-based explanations of CNN, such as Integrated Gradients (IG) and Guided Backpropagation, showed low robustness on neural network models trained on the ImageNET dataset. However, the authors do not provide empirical or theoretical evidence that the retrained model will have the same properties as the original model we intend to explain. In addition, studies have shown that the correlation relationship among features does not hold in the new model after the retraining step (Nguyen and Martínez 2020). In Agarwal et al. (2022), the authors propose the OpenXAI framework that includes numerous robustness measures. They showed that model-based gradient explanation techniques such as Gradient∗Input (Shrikumar et al. 2016) provided more robust explanations than LIME and SHAP explanations across numerous datasets.

2.2.2 Ground truth from synthetic datasets

Some studies have proposed to evaluate explanations directly based on extracting ground truth importance scores from synthetic datasets. These studies aim to tackle the first limitation of the robustness measure, as discussed in the previous section. The core assumption behind these evaluation methods is that obtaining ground truth from the black-box models on arbitrary data is challenging. Therefore, we can simplify the data these models are trained on. Using specific data generation processes enables these methods to control the importance of each feature for the generated labels *prior* to the training phase of explained models. Local explanations that provide feature importance scores similar to these priors are considered the most accurate.

The SenecaRC algorithm (Guidotti 2021) generates data from a polynomial function that can include varying operators such as *sin* or *cos* in its polynomial terms. After that, a sample is generated based on the chosen polynomial function. Lastly, the algorithm returns the ground truth importance scores for the explained instance x

based on the following steps: (1) the closest instance $x*$ to x on the decision boundary of an explained model, g, is found, and (2) the derivative of the ground truth polynomial is evaluated at this point and returned as true importance scores for x.

In Liu et al. (2021), the authors provide a set of synthetic datasets and evaluate the quality of local explanations using (robustness) measures such as faithfulness and fidelity. SHAP and SHAPR (Aas et al. 2021) explanations were observed to have higher faithfulness compared to LIME and Model Agnostic SuPervised Local Explanations (MAPLE) explanations for the considered set of synthetic datasets. In their evaluation, the authors show that LIME, SHAP, and MAPLE (Plumb et al. 2018) explanations fail to provide accurate explanations for (synthetic) tabular datasets with large numbers of uninformative features.

In Agarwal et al. (2022), the authors proposed a synthetic SynthGauss dataset. They argue that their proposed dataset is more suitable for evaluating explanations than the dataset in Liu et al. (2021) since features are independent in their proposed dataset and local neighborhoods do not overlap in the dataset. In their study, model-based gradient explanations such as SmoothGrad (Omeiza et al. 2019)was observed to outperform LIME and SHAP explanations across numerous datasets.

The main limitations of evaluation approaches based on synthetic ground truth are two-fold: (1) Since the priors of feature importance scores are set before the explained model is trained, there are no guarantees that the model has learned the relationship between features and the label in the synthetic dataset according to our prior importance scores (Faber et al. 2021) (2) Synthetic datasets are not complex in terms of empirical feature distribution and interactions between their features unlike many tabular datasets (Guidotti 2021). As a result, we cannot directly conclude that since a local explanation is inaccurate on these synthetic datasets, it is also inaccurate on larger and more complex datasets.

2.2.3 Ground truth using interpretable models

As mentioned earlier, we cannot directly extract ground truth importance scores from complex black-box models. The extraction of the ground truth can be made easier if we explain a simpler class of machine learning models. The methods that extract ground truth importance scores from interpretable models follow this assumption. The strength of these evaluation methods is that we are no longer restricted to evaluating local explanations on simplified datasets. These methods obtain the ground truth importance scores extracted directly from the trained model. Unlike the ground truth from synthetic datasets, we can guarantee that these importance scores are directly extracted from the knowledge that exists in the trained model. However, this comes at the cost of only being able to evaluate simple models. This type of evaluation can thus only be used as a sanity check and not to evaluate the accuracy of any model on any dataset.

In Agarwal et al. (2022), the authors proposed to extract ground truth importance scores from the weights of Logistic Regression. Based on their evaluation, model-based explanations such as SmoothGrad have larger similarities to their proposed ground truth than LIME and SHAP explanations. The main limitation of their baseline for extracting ground truth is that the authors have used the weights of Logistic

Regression, a *global* explanation, as their baseline for evaluating *local* explanations. Global explanations are one vector of the feature importance scores for an entire dataset that is equal for all instances (Freitas 2014). On the other hand, local explanations exhibit properties of the locality of that instance in the data input space (Ribeiro et al. 2016). Based on this, measuring the similarity of all unique local explanations for each instance to the global explanation can lead to incorrect conclusions. [5].

In this study we propose a method that extracts the local ground truth importance scores for three linear additive models, linear and logistic regression and Naive Bayes. In contrast to the aforementioned method, we thus know how much each feature contributes to the model's predicted output and can directly compare the importance scores generated by a local explanation technique.

3 Motivating example

In this section, we show an example that highlights reasons the current evaluation methods cannot provide the correct evaluation method for evaluating local explanations of linear additive models. Let us reiterate that we expect a local ground truth importance score to include some of the instance's locality in the model's decision space. We show that the currently available approaches either allocate equal ground truth measures to all instances, disregarding the instance locality, or fail to allocate the correct importance to all features. Our example uses Seneca-RC's synthetic dataset generation and compares the baselines from synthetic datasets proposed by Guidotti (2021), the ground truth proposed by Agarwal et al. (2022) and robustness measures (Hsieh et al. 2020; Fong and Vedaldi 2017).

Let $Y = 2x_0 - x_1$ be the data generation process where features x_0 contribute positively and x_1 negatively to the label. We sample one thousand instances from Seneca-RC's data generation process where no extra redundant features are added, and we set the noise level to 0.3. We train a Logistic Regression model on this generated dataset[6]. The model achieves a test accuracy of 0.98 on this dataset. The decision boundary (see Fig. 1) shows that the model has correctly identified that both features in combination are important for separating instances from different classes. The arrows on the top of each instance represent the ground truth importance scores based on each evaluation method.

The Seneca-RC ground truth importance scores are all equal to the vector, $[1, -1]$, irrespective of the position of the instance in the prediction space or the decision boundary of the model. This is because the derivative of the data generation process with respect to each feature is a constant value. Therefore, the ground truth from

[5] The global explanation of Logistic Regression as a benchmark for evaluation is not mentioned directly in the study of Agarwal et al. (2022) . However, it can be seen in the https://github.com/AI4LIFE-GROUP/OpenXAI/blob/a335201e4f9f4ddad97f8b1a0f6ff9fe750903bf/openxai/ML_Models/LR/model.pyin the code repository released with this study

[6] Since (Agarwal et al. 2022) has only provided the ground truth importance scores for Logistic Regression, we provide an example with a Logistic Regression model.

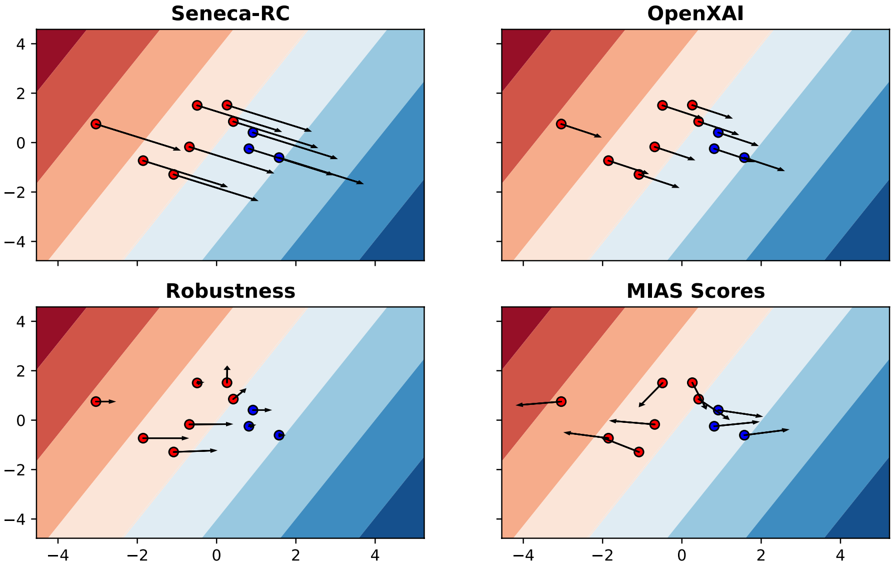


Fig. 1 Comparison of ground truth importance scores of Seneca-RC and OpenXAI along with robustness values of each feature compared to our Model-intrinsic Additive Score (MIAS). The dataset is generated by the Seneca-RC algorithm and a Logistic Regression model. The ground truth importance score for each instance is visualized as vectors on the top of each instance

Seneca-RC does not reflect the true locality of instances in the decision space of the trained Logistic Regression model (Sect. 2.2.2).

The ground truth of OpenXAI (Agarwal et al. 2022) is also constant across all instances. This is because the Logistic Regression model weights are directly used as the baseline for obtaining ground truth importance scores for all instances in this approach. Since the model weights are the summary of the importance of features for all instances, all instances are then evaluated using one equal ground truth scores regardless of their position in the decision space (see Sect. 2.2.3 for details).

For the robustness measures, we no longer show the ground truth but the robustness values of each feature for every instance. The value of the arrow on top of instances shows the absolute change in the predicted scores of class one (blue circles) after that feature is nullified separately. We nullify each feature using the average values of that feature in the dataset as it is generally practiced in tabular datasets (Liu et al. 2021; Montavon et al. 2018; Molnar et al. 2022) . We can see that for most instances, robustness measures do not set any importance to the second feature on the y-axis even though it plays an important role in the linear boundary of logistic regression and the data generation process. Moreover, an instance will receive zero robustness by default along an axis, i.e. for a feature, if its feature values are similar to the empirical average of each feature. This is because nullifying those features will not affect the predicted output.

On the other hand, our proposed Model-Intrinsic Additive Scores (MIAS) allocate different values for instances that are located on the decision plane of the Logistic Regression model. As we show later in Sect. 4.3, the MIAS score of Logistic

Regression models sets importance to both features in explaining the log odds ratio of the model. We can also see that the instances will then have arrows toward the subspace with maximum log odds of their predicted class visualized by the shades in the background. We can see that the MIAS vectors of instances close to the decision boundary are more different since the uncertainty in the model's predicted output is larger in those parts of the plane. In the next Section, we present how we can calculate the MIAS scores of linear additive models such as Linear and Logistic Regression and Gaussian Naive Bayes.

4 Evaluation methodology

In this section, we introduce our proposed evaluation framework in detail. As shown in Sect. 3, all current evaluation measures have shortcomings in the way in which ground truth importance scores are allocated for linear additive models.

Our study proposes a new method for evaluating local model-agnostic explanations of linear additive models. Our evaluation methods fall into the category of *evaluation methods using interpretable models* (Sect. 2.2.3). Unlike the work of Agarwal et al. (2022), we follow a more principled method for the evaluation of local explanations. We extract the ground truth by extracting individual additive terms from the prediction function of any class of linear additive models, e.g., Logistic and Linear Regression and Naive Bayes.

Our approach can extract ground truth importance scores from models where the prediction function is linear additive, e.g. in Linear regression models. Moreover, we also extract the ground truth for models in cases where the prediction function is not linear additive directly but can be transformed into a linear additive function such as in Logistic Regression and Naive Bayes models. As shown in Sect. 3, our ground truth importance scores allocate the ground truth on a single instance level.

In Sect. 4.1, we discuss the main logic behind our evaluation method to extract our so-called Model-intrinsic Additive Score (MIAS) for linear additive models such as Linear and Logistic Regression and Gaussian Naive Bayes. Lastly, we argue for the choice of similarity metric in Sect. 4.5.

4.1 Model-intrinsic additive scores

As we mentioned, we follow a more principled approach to extract our ground truth importance scores.We formulate the problem as follows. In Eq. 1, we can see that we have one linear additive decomposition of $f(x)$. If the prediction function f can also be represented as an additive sum similar to Eq. 1 like the following:

$$f(x) = \sum_{j=1}^{M} \lambda_j x_j, \tag{4}$$

we can measure the explanation accuracy by measuring the similarity of individual additive terms $\phi_j x_j$, importance scores of feature j in Eq. 1, to $\lambda_j x_j$. This is possible as both equations are linear additive decompositions of $f(x)$. An additive structure like Eq. 4 is directly visible in linear additive models such as linear regression and can also be extracted in Logistic Regression and Naive Bayes models. Even though these additive structures have long existed in the machine learning literature, they have to the best of our knowledge, not been used as a means to evaluate local explanations.

Definition 1 Local Explanation accuracy: Let Φ be a local explanation for instance x. The local explanation accuracy is defined as $\sum_{j=1,\ldots,M} d(\phi_j x_j, \lambda_j x_j)$ where λ_j is the weight for feature j in form of 4 and d is a similarity metric. Based on this, we call $\lambda_j x_j$ a Model-Intrinsic Additive Score (MIAS) for feature j.

We want to highlight that our proposed MIAS score includes the input instance's feature value in calculating the ground truth. In other words, unlike global explanations, each MIAS score is specific to the single instance explained. The inclusion of feature values in calculating our ground truth is similar to the proposal of Liu et al. (2019) in which the input feature values are used for obtaining the local gradient-based explanations for neural network models.

Algorithm 1 summarizes the logic of our evaluation framework. On a high level, to evaluate an explanation technique g of the linear additive model f, we extract the Model-intrinsic Additive Scores (MIAS) Λ. After that, we obtain a local model-agnostic explanation Φ for a single instance x_n from the explanation technique g. We then compute the similarity between Λ and Φ using similarity metric ρ, i.e. $rho(||\Lambda, \Phi||)$.

In general, we are interested in comparing explanation accuracy across different datasets. Therefore, we run Algorithm 1 over the test sets of each dataset. The higher the average values of $r_{x_n f, g}$ are over a test set, the more accurate the explanations of g when explaining model f are for that dataset.

```
Algorithm 1 Evaluating explanation
Input $x_n$: instance
Functions $f$: white-box model, $g$: explanation method, $t$: function to extract MIAS importance scores
          $\rho$: similarity measure
Output $r_{x_n,f,g}$: similarity value
$\Lambda \leftarrow t(f, x_n)$
$\Phi \leftarrow g(f, x_n)$
$r_{x_n,f,g} \leftarrow \rho(\Lambda, \Phi)$
```

The logic behind the extraction of MIAS importance scores for Linear Regression, Logistic Regression, and Gaussian Naive Bayes models are discussed in Sects. 4.2, 4.3 and 4.4 respectively (function t in Algorithm 1).

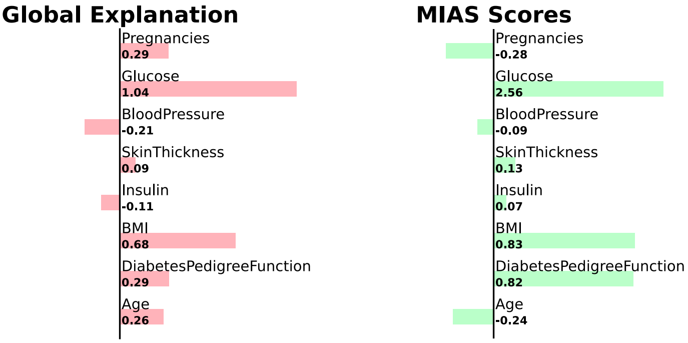


Fig. 2 (Left): The global explanations of the Logistic Regression model trained on the Pima Indians dataset, which is equal for all test instances if used as a local ground truth importance score. (Right) MIAS scores of a test instance of the same dataset. MIAS scores differ for each test instance depending on their feature values

4.2 Linear regression

As we said earlier, the linear regression model has a linear additive structure in the following form:

$$f(x) = w_0 + \sum_{j=1}^{M} w_j x_j \tag{5}$$

where w_j is the weight for feature j and w_0 represents the intercept and x_j is the j-th component of x. In our study, we consider $w_j x_j$ as the MIAS score for the contribution of feature j to the predicted output of $f(x)$, i.e. $\Lambda = w_j x_j$.

4.3 Logistic regression

Given weights $w \in \mathbb{R}^{M+1}$ and an instance $x_n \in \mathbb{R}^M$, a logistic regression model is defined as:

$$P(y_n = c \mid x_n, w) = \frac{1}{1 + e^{-\sum_{m=0}^{M} w^m x_n^m}} \tag{6}$$

where $x_n^0 = 1$. Even though there is no direct linear additive form of this prediction function, we can derive an additive decomposition of a model prediction using the log odds ratio for x_n concerning class $c \in \{0, 1\}$:

$$log \frac{P(y_n = c \mid x_n, w)}{P(y_n = \neg c \mid x_n, w)} = \sum_{m=0}^{M} w^m x_n^m \tag{7}$$

where $\neg c$ is the complement of class c and $\lambda_n^m = w^m x_n^m$ is the Model-Intrinsic Additive Score (MIAS) for feature m. Note that in this case, we explain the log odds and therefore, $f(x) \leftarrow \log \frac{P(y_n = c \mid x_n, w)}{P(y_n = \neg c \mid x_n, w)}$ in Eq. 4.

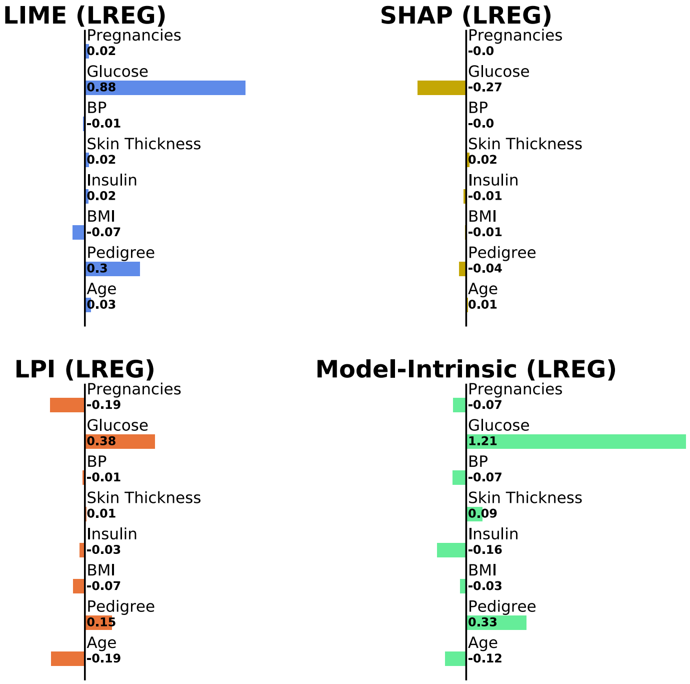


Fig. 3 The feature importance scores of MIAS as well as LIME, SHAP, and LPI explanations for a single instance from the Pima Indians data set when explaining a logistic regression prediction

In Fig. 2, we compare the weights of a Logistic Regression model (its global explanations) to the MIAS scores obtained for a single test instance of the Pima Indians dataset. Notice that the global explanation will be the same for all test instances, whereas MIAS scores are different for each instance.

4.4 Naive Bayes

Given input $x_n = (x_n^1, \dots, x_n^M)$ and a mean and variance vector, $\mu_c \in \mathbb{R}^M$ and $\sigma_c \in \mathbb{R}^M$, we can apply the Bayes theorem:

$$P(y_n = c \mid x_n) = \frac{P(x_n \mid y_n = c)P(y_n = c)}{P(x_n)}, \tag{8}$$

where the likelihood $P(x_n \mid y_n = c)$, under the naive assumption of conditional independence, can be computed as:

$$\prod_{m=1}^{M} P(x_n^m \mid y_n = c) = \prod_{m=1}^{M} \mathcal{N}(x_n^m \mid \mu_c^m, \sigma_c^m). \tag{9}$$

Similar to the case of logistic regression, the prediction function does not naturally decompose into additive parts. However, the log odds ratio for an instance x_n for class c has an intrinsic natural additive decomposition:

$$\log \frac{P(y_n = c \mid x_n)}{P(y_n = \neg c \mid x_n)} = \sum_{m=1}^{M} \log \frac{\mathcal{N}(x_n^m \mid \mu_c^m, \sigma_c^m)}{\mathcal{N}(x_n^m \mid \mu_{\neg c}^m, \sigma_{\neg c}^m)} + const. \tag{10}$$

where $const = \log \frac{P(y_n = c)}{P(y_n = \neg c)}$. Based on this, the MIAS importance scores of feature m is $\lambda_n^m = \log \frac{\mathcal{N}(x_n^m \mid \mu_c^m, \sigma_c^m)}{\mathcal{N}(x_n^m \mid \mu_{\neg c}^m, \sigma_{\neg c}^m)}$. Note that in this case, instead of $f(x)$, we explain the log odds ratio prediction in Eq. 4.

In Fig. 3, an example of our MIAS scores for a single instance is visualized in comparison to explanations of LIME, SHAP, and LPI for the Pima Indians dataset for the Logistic Regression model. See the appendix for a similar visualization for the Naive Bayes model on this dataset.

4.5 Similarity measure

We measure accuracy in terms of how similar an explanation is to the MIAS importance scores. Several studies have used measures such as Cosine or Euclidean distance (Montavon et al. 2018; Yang and Kim 2019) to measure the similarity of explanations.

Similar to Ghorbani et al. (2019), we argue that the Spearman's Rank correlation sometimes may be a more suitable measure for comparing explanations in tabular datasets, as it is not affected by the absolute values of importance scores but only the ranking of these values. Additionally, the interpretation of explanations might differ across different types of explanation techniques. The rank correlation measure makes it possible to compare feature importance scores between additive and non-additive explanations techniques that cannot be directly compared. Lastly, in contrast to Euclidean and Cosine similarity, the metric comes with interpretable measures of direction and strength. One drawback of using a rank-based measure is that it might be sensitive if a dataset has many unimportant feature dimensions. In this case, the performance across all explanation techniques will be low as the ranking of unimportant features will vary randomly.

An incorrect choice of a similarity metric that does not fit the use case may lead to wrong conclusions. To illustrate this, we provide an example comparing two local explanations using Euclidean and Cosine similarity along with Spearman's rank correlation. Suppose we need to measure the accuracy of two different explanations $\phi_1 = [0.21, 0.1, 0.32]$ and $\phi_2 = [0.21, 0.3, 0.12]$ to the ground truth score $\lambda = [0.32, 0.2, 0.42]$.

$$\begin{aligned} \mathit{Euclidean}\, S(\lambda, \phi_1) &= 0.179 & \qquad \mathit{Euclidean}\, S(\lambda, \phi_2) &= 0.28 \\ \mathit{Spearman}\, C(\lambda, \phi_1) &= 1 & \qquad \mathit{Spearman}\, C(\lambda, \phi_2) &= -1 \\ \mathit{Cosine}\, S(\lambda, \phi_1) &= 0.99 & \qquad \mathit{Cosine}\, S(\lambda, \phi_2) &= 0.81 \end{aligned}$$

Based on Spearman's rank correlation, the ranking of ϕ_1 correlates perfectly with λ, while the ranking of ϕ_2 negatively correlates with λ. Using this rank-based metric, we can thus conclude that explanation ϕ_1 is more accurate than ϕ_2. The Euclidean,[7] and Cosine Similarity instead vote in favor of ϕ_2 as the more accurate explanation. We show the role of similarity metric in our experiments later in Sect. 5.2.6.

5 Empirical investigation

In this section, we will present the results of our empirical investigation. We describe the experimental setup in Sect. 5.1. After that, we provide in Sect. 5.2 the result of our empirical experiments on local explanation accuracy for all explanation techniques and models considered.[8]

5.1 Experimental setup

In this section, we describe our experimental setup for the dataset and models that we used for obtaining explanations in Section 5.1.1. After that, we provide some information about the hyperparameters for generating explanations in Sect. 5.1.2.

5.1.1 Data and model

We assess the proposed evaluation framework using the total of 40 different tabular datasets concerning both (binary and multi-class) classification and regression tasks. All the datasets are publicly available at the UCI, Kaggle, or Keel repositories.[9] Unless otherwise stated, the numerical features are standardized and categorical features are one-hot encoded. For datasets for which no separate test set has been provided at the source, a random hold-out set of 25% was used. The information for each dataset is shown in the appendix.

We trained logistic and linear Regression along with Gaussian naive Bayes models using the aforementioned datasets. To tune the hyper-parameters of the logistic regression models, grid-search was employed. Hyper-parameters were chosen after 100 trials with the hyper-parameter space consisting of L1 and L2 regularization with the regularization parameter selected from a grid of values between 0 to 4. Tables 1 and 2 report the test accuracy of the models.

[7] In our study, we define Euclidean similarity as $1/(\epsilon + d)$ where d is the Euclidean distance and ϵ is the machine epsilon of Python.

[8] The code for experiments is available at: https://github.com/amir-rahnama/can_local_explanations_explain_lam.

[9] See appendix for more information.

Table 1 Information about the datasets used in our study for classification tasks

Dataset	Numerical	Categorical	Total	LOGR	NB
Adult	6	8	14	0.85	0.56
Attrition	16	17	33	1	1
Audit	23	3	26	0.99	0.96
Banking	5	10	15	0.91	0.83
Banknote	4	0	4	0.98	0.85
Breast cancer	30	0	30	0.97	0.92
Churn	8	2	10	0.81	0.8
Donor	42	6	48	1	1
Haberman	3	0	3	0.66	0.65
Hattrick	20	4	24	1	0.79
Heart disease	6	7	13	0.83	0.8
Hr	2	10	12	0.78	0.38
Insurance	4	6	10	0.99	0.07
Iris	4	0	4	1	1
Loan	5	8	13	1	1
Pima Indians	8	0	8	0.8	0.77
Seismic	14	4	18	0.95	0.42
Spambase	58	0	58	0.92	0.81
Thera	10	2	12	1	1
Titanic	6	2	8	0.79	0.79

The number of numerical, categorical, and total number of features and the test accuracy for Logistic Regression *LOGR* and Naive Bayes *NB* models are presented for all our classification datasets

5.1.2 Generating explanations

For LIME and SHAP, the official Python packages TabularLIME (Ribeiro et al. 2016) and KernelShap (Lundberg and Lee 2017) have been used. We want to emphasize that the KernelShap explainer is model-agnostic, and it outputs approximated SHAP values. This contrasts with model-based explainers such as LinearSHAP where the SHAP values are analytically deducible from closed-form equations (see Lundberg and Lee 2017 for details). In our study, we are comparing model-agnostic explanations where the explainers make no assumptions on the class of machine learning models they are explaining. The number of samples generated for LIME and SHAP is 5000. We show the logic behind choosing this sample size in Sect. 5.2.4. The sample size of LPI is equal to the size of the training set as suggested in Casalicchio et al. (2018). This means the sample size of LPI is significantly smaller than the size of LIME and SHAP on average.

As mentioned earlier in Sect. 4.1, given that the MIAS scores are extracted from the log odds ratios of instances for logistic regression and naive Bayes, we need to pass in the log odds ratio prediction function to all explanation techniques. For this, all we need to do is to write the log odds prediction function and

Table 2 The total number of features and the test set mean squared error for the linear regression (LR) model for all our regression datasets

Dataset	Features	Mean squared error
Anacalt	7	0.633
Bank8Fh	8	0.01
Bank8Fm	8	0
Bank8Nh	8	0
Bank8Nm	8	0
Delta A	5	0.4
Delta E	6	0.035
Istanbul	7	0
Kin8Fm	8	0
Kin8Nh	8	0.05
Kin8Fh	8	0
Kin8Nm	8	0
Mortgage	15	0.163
Puma8Fh	8	10.53
Puma8Fm	8	1.67
Quakes	3	0.041
Treasury	15	0
Wine red	11	0.013
Wine white	11	0.053
Wizmir	9	1.23

pass that to explanation techniques. This is possible as, in both LIME and SHAP packages, one can pass in any desired prediction function for obtaining explanations. In the case of LPI, we have replicated the algorithm proposed in Casalicchio et al. (2018) such that the importance scores are calculated based on the difference in the predicted log odds ratio scores instead of the prediction function following the permutation of each feature for the case of logistic regression and naive Bayes models (see Sect. 2.1). Lastly, for each instance, our explanations are obtained for the predicted class by the explained classification model. To focus on evaluating the important features, we compare the absolute values of importance scores from local explanations with our proposed MIAS scores as it is common in the tabular datasets (Ribeiro et al. 2016; Lundberg and Lee 2017).

5.2 Experiments

In this section, we provide the result of measuring the explanation accuracy for all of our studied explained models, namely linear and logistic regression and naive Bayes. Discussions about factors contributing to the average accuracy values are presented separately in Sects. 5.2.2 to 5.2.5. The study of these factors is based on the explanation accuracy when using Spearman's rank correlation. We discuss the effect of the

Table 3 Average explanation accuracy based on Spearman's rank correlation for LIME, SHAP (additive), and LPI (non-additive) explanations when explaining linear regression model

Dataset	LIME	SHAP	LPI
Anacalt	0.999	**1**	0.424
Bank8Fh	0.892	**0.952**	0.858
Bank8Fm	0.891	**0.952**	0.843
Bank8Nh	0.824	**0.866**	0.6
Bank8Nm	0.835	**0.859**	0.63
Delta A	**0.928**	0.908	0.5
Delta E	0.955	**0.971**	0.627
Istanbul	0.876	0.899	**0.956**
Kin8Fh	0.929	0.926	**0.997**
Kin8Fm	0.943	0.943	**0.994**
Kin8Nh	0.901	0.92	**0.997**
Kin8Nm	0.883	0.915	**0.995**
Mortgage	0.955	**0.972**	0.614
Puma8Fh	0.783	0.93	**0.955**
Puma8Fm	0.747	**0.955**	0.951
Quakes	0.825	**0.845**	−0.105
Treasury	0.886	**0.946**	0.656
Wine red	0.864	**0.881**	0.143
Wine white	**0.872**	0.867	0.261
Wizmir	0.898	**0.897**	0.643
Average	0.885	0.92	0.677
Standard deviation	0.059	0.041	0.306

Bold values indicate the explanation technique with the highest average explanation accuracy

choice of similarity metrics on the explanation accuracy in Sect. 5.2.6. The effect of the empirical distribution of features on the explanation accuracy of classification models is discussed in Sect. 5.2.7. In Sect. 5.2.8, we measure the robustness of the explanations on our datasets and whether explanations with large average accuracy are the most robust.

5.2.1 All datasets

We first investigate explanation accuracy of local explanations for linear regression models. In Table 3, the average explanation accuracy of additive (LIME and SHAP) and non-additive (LPI) explanations of Linear Regression models are shown. The explanation accuracy is the similarity of each explanation to our proposed MIAS score based on Spearman's rank correlation. Overall, SHAP has a larger average explanation across all regression datasets than LIME and LPI. LPI outperforms other techniques in Istanbul and Kin8Nm datasets and LIME in Wine White and Delta A.

Table 4 Average explanation accuracy for LIME, SHAP and LPI explanations when explaining Logistic Regression and naïve Bayes Models based on Spearman's rank correlation

Model →	Logistic regression			Naïve Bayes		
Dataset	LIME	SHAP	LPI	LIME	SHAP	LPI
Adult	−0.086	−0.061	**0.227**	−0.427	**0.013**	−0.548
Attrition	−0.005	0.001	**0.049**	−0.001	**0.261**	0.254
Audit	**0.075**	−0.004	0.029	−0.099	**0.061**	−0.269
Banking	−0.071	−0.093	**0.007**	0.025	0.122	**0.328**
Banknote	**0.918**	0.9	0.778	0.844	0.678	**0.904**
Breast Cancer	**0.882**	0.871	0.803	**0.753**	0.722	0.455
Churn	0.121	**0.146**	−0.124	0.174	**0.251**	0.012
Donor	0.128	−0.071	**0.163**	**0.021**	−0.221	−0.221
Haberman	0.695	0.461	**0.708**	0.786	−0.026	**0.877**
Hattrick	−0.039	**0.014**	−0.046	−0.111	**−0.013**	−0.438
Heart Disease	0.059	**0.242**	0.152	−0.117	**0.343**	−0.396
Hr	0.095	0.149	**0.386**	−0.24	**−0.081**	−0.28
Insurance	−0.257	**−0.172**	−0.201	−0.525	**−0.23**	−0.494
Iris	0.832	0.776	**0.848**	0.872	**0.877**	0.78
Loan	**0.487**	−0.22	0.463	**0.287**	0.176	0.178
Pima Indians	0.841	**0.863**	0.593	**0.738**	0.541	0.606
Seismic	**−0.197**	−0.236	−0.402	0.127	**0.14**	−0.187
Spambase	**0.856**	0.223	0.552	−0.389	0.179	**0.463**
Thera	−0.163	**0.245**	−0.199	0.256	**0.481**	**0.481**
Titanic	−0.017	**0.214**	−0.043	−0.236	**0.293**	−0.196
Average	**0.258**	0.212	0.237	0.137	**0.228**	0.115
Standard Deviation	0.409	0.363	0.369	0.435	0.298	0.465

Bold values indicate the explanation technique with the largest average accuracy

Since our proposed similarity measure is an average correlation, we expect the average accuracy values to be significant, e.g., above 0.7 and not lower than 0.5 (Ross 2017) . The average explanation accuracy of LIME and SHAP explanations passes this threshold across all datasets. Due to the consistent behavior of these explanations on regression datasets, we can consider these explanations accurate for explaining linear regression models. Surprisingly in some datasets such as Istanbul, Kin8Fh, and King8FM, LPI provides the largest average explanation accuracy compared to LIME and SHAP. However, this trend is inconsistent for other datasets, for example, in Wine White and Red or Quakes. We raised a question in Sect. 1 about whether the linear additivity of local explanation can be an advantage in providing accurate local explanations. Our results suggest that the additivity of explanations is indeed advantageous when explaining the linear regression model. In Sect. 5.2.5, we show that one main reason behind LPI's low average explanation accuracy is the large variance in its accuracy values.

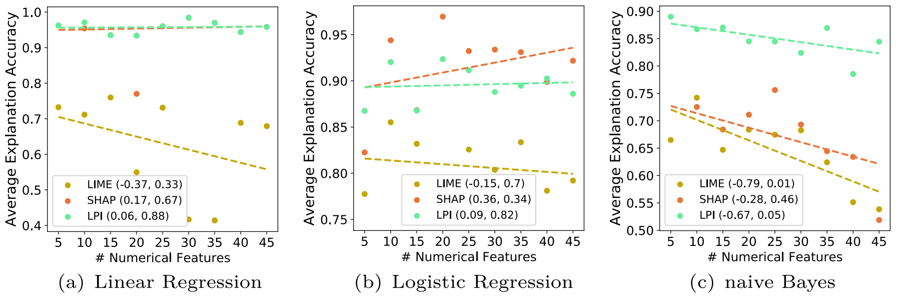
The result of average explanation accuracy for Logistic Regression and Naive Bayes models are shown in Table 4. LIME provides the largest average accuracy when explaining Logistic Regression, whereas SHAP explanations have the largest average accuracy when explaining the Gaussian Naive Bayes model. The difference between the average explanation accuracy over all classification datasets is lower when explaining the Naive Bayes models than the Logistic Regression. To our surprise, LPI outperforms the additive explanations of LIME and SHAP across numerous datasets, e.g., Donor and Haberman when explaining Logistic Regression and Banknote and Spambase when explaining naive Bayes models. Unlike our results for the explanations of Linear Regression models, we can see that the average accuracy of local explanations can be significantly low across numerous datasets such as Adult, Attrition, Audit, Churn, Donor, Hattrick, Hear Disease, HR, Insurance, Seimsimc, Thera, and Titanic. For example, the largest average explanation accuracy for the Audit dataset is obtained by LIME, with 0.075 for Logistic Regression, and SHAP, with 0.061 for Naive Bayes. This means that even the best-performing explanations could not find the correct ranking of the most important features, even for 10% of instances in the test dataset. Our results suggest that our study's explanations of linear additive classification models do not exhibit acceptable accuracy to pass our sanity check.[10]

In some datasets, explanation accuracy reaches an acceptable threshold, e.g., for all explanations of both models in the Banknote and Iris and Pima Indians dataset, where all average explanation accuracy values are above 0.7. This is partly because this dataset has few numerical and no categorical features. In contrast, the low values of the explanation accuracy of the Donors dataset can be partially explained by the existence of large number of categorical features. We will discuss the effect of data on explanation accuracy later in Sect. 5.2.2.

When explaining the Logistic Regression models, the low average explanation accuracy values in HR and Titanic datasets for all explanation techniques can be explained by the low model predictive performance on the test set, i.e., poor model generalization. We discuss the effect of model generalization further in Sect. 5.2.3

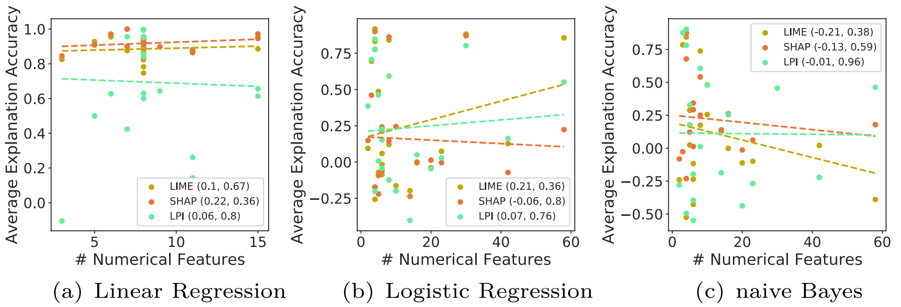
However, there are cases where we cannot blame the model's generalization as the main factor behind the low values of explanations accuracy. For example, Logistic Regression and Naive Bayes models achieve an acceptable generalization accuracy on the Thera and Heart Disease datasets, respectively. Yet, the average explanation accuracy in these datasets is relatively low across all explanation techniques. In Sect. 5.2.5, we show that low average explanation accuracy in these classification datasets is caused by the large standard deviation of explanation accuracy within each dataset. In the presence of a large standard deviation in the explanation accuracy, although explanations can be very accurate for a subset of instances, they are also inaccurate for others.

[10] See the appendix for a study of the statistical significance of the average explanation accuracy for all explanation techniques for all linear additive models.



(a) Linear Regression (b) Logistic Regression (c) naive Bayes

Fig. 4 The average explanation accuracy as the number of numerical features increases in synthetic datasets. Pearson Correlation values, together with the p-values, are included in the legend



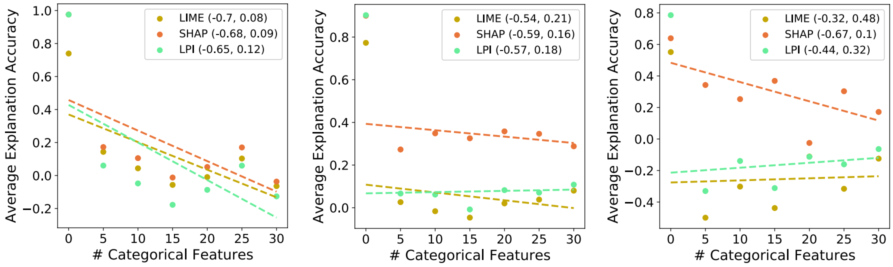
(a) Linear Regression (b) Logistic Regression (c) naive Bayes

Fig. 5 Linear relationship between average explanation accuracy and the number of numerical features over all tabular datasets. Pearson Correlation values, together with the p-values, are included in the legend

5.2.2 The data effect

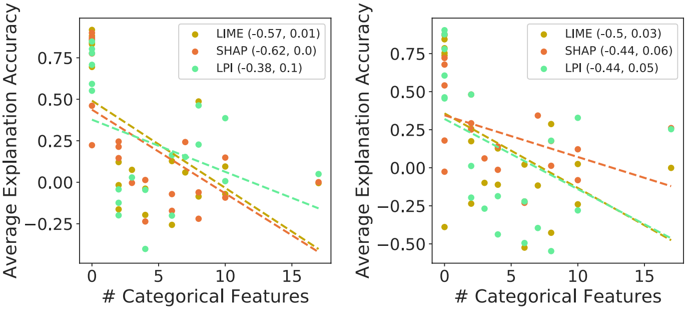
In some studies (Molnar et al. 2022; Guidotti 2021), the authors have provided specific synthetic datasets in which the accuracy of local model-agnostic explanations is worsened with an increase of the number of numerical and categorical features. In this section, we investigate whether there is a linear relationship between the number of numerical, categorical, and pairwise correlated features and the average explanation accuracy at the dataset level. We first investigate this in synthetic cases and then in our tabular datasets. Overall, we show that the linear relationship between the data-related factors highly depends on the type of linear additive models and the explanation techniques used for obtaining explanations.

5.2.2.1 Numerical Features Let us begin with studying the effect of numerical features on explanation accuracy in synthetic datasets. We use ScikitLearn (Kramer and Kramer 2016) 's classification and synthetic regression dataset generator for the synthetic datasets. We considered 20% of all features as uninformative. In the experiment, we increase the number of numerical features in our synthetic datasets from



(a) Linear Regression (b) Logistic Regression (c) naive Bayes

Fig. 6 The change in average explanation accuracy and the number of categorical features. In these datasets, K number of features are transformed into categorical features, and the rest of $40 - K$ features are fixed as before. Pearson Correlation values, together with the p-values, are included in the legend

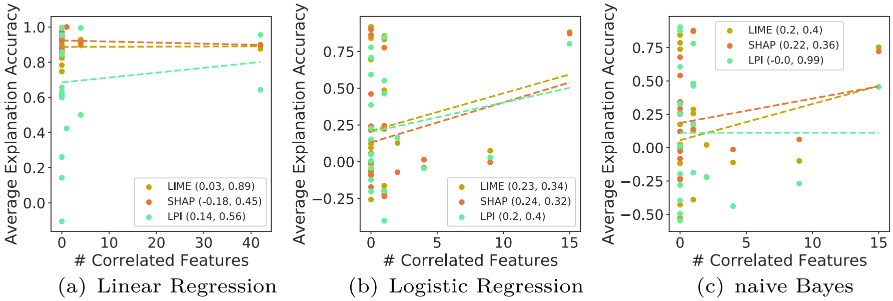


(a) Logistic Regression (b) Naive Bayes

Fig. 7 linear relationship between average explanation accuracy and the number of categorical features across all classification datasets. Pearson Correlation values, together with the *p*-values, are included in the legend

1 to 45 without the presence of any categorical features. To control for the effect on model generalization, we have only considered models with relatively similar accuracy values (See Table 12 in Appendix). Figure 4 shows that increasing the number of numerical features minimally affects the average explanation accuracy of LIME and SHAP explanations of Linear Regression models, yet it decrease the average accuracy of LPI explanations for this model. The SHAP (LIME) explanations of Logistic Regression have larger (smaller) average accuracy as the number of numerical features increases. No significant change in the average accuracy of LPI explanations are visible in this model. For Naive Bayes explanations, the average accuracy of all explanations decreases with an increase in the number of numerical features.

We investigate whether the same trends hold in our tabular dataset. In Fig. 5 , we can see the effect of these factors for the explanations of all linear additive models. In this figure, each point represents the average explanation accuracy of a single tabular dataset. The figure shows that numerical features have minimally positive effects on the average explanation accuracy of Linear Regression models. With an increased number of numerical features, the average explanation accuracy of LIME



(a) Linear Regression (b) Logistic Regression (c) naive Bayes

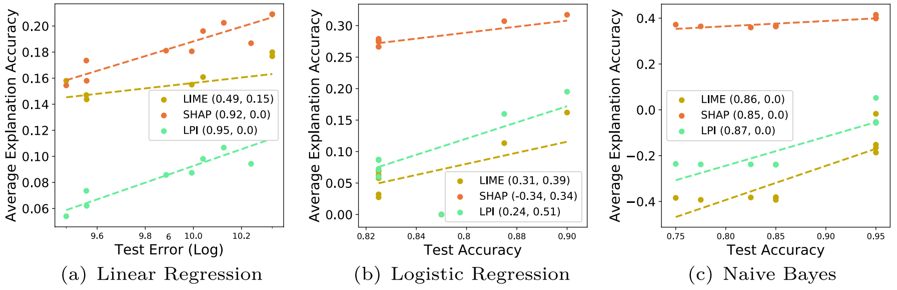
Fig. 8 The change in average explanation accuracy with the number of pairwise correlated features in tabular datasets. Pearson Correlation values, together with the *p*-values, are included in the legend

for Logistic Regression increases, and the average accuracy of LIME and SHAP decreases for Naive Bayes explanations. Some trends are similar between the synthetic and tabular datasets. LPI and SHAP explanations of Linear Regression models show minimal change in accuracy with an increase of numerical features. LPI shows the same trend for the explanations of Logistic Regression. Lastly, the average accuracy of LIME and SHAP explanations of Naive Bayes models decreases with an increase in the number of numerical features.

5.2.2.2 Categorical features To analyze the effect of categorical features, we start with a synthetic setting. We fix the total number of features in our synthetic dataset generator to 40 numerical features. In each step, we increasingly turn K of these features into categorical features and then transform these features using one-hot encoded categorical ones. Meanwhile, we keep the $40 - K$ features unchanged. The number of categorical features is calculated before they are one-hot encoded. In Fig. 6, the average explanation accuracy of all explanations of Linear Regression model decrease as the number of categorical features increase. The average accuracy accuracy of LIME and SHAP explanations of Logistic Regression shows a slight decrease as the number of categorical features increases. Lastly, the average accuracy of SHAP explanations of naive Bayes models decreases with an increase in the number of numerical features.

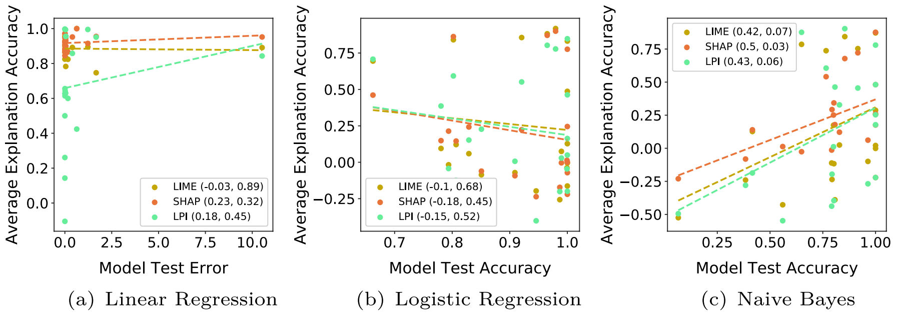
In Fig. 7, we perform similar analyses to study the effect of categorical features in our tabular classification dataset. With an increased number of categorical features, all explanations of Logistic Regression and the Naive Bayes model show a steady decrease in their average accuracy. We can see some similar trends between the synthetic and tabular datasets. LIME and SHAP explanations of Logistic Regression, and SHAP explanations of naive Bayes models show a decrease in average accuracy when the number of categorical features increases.

5.2.2.3 Correlated features Lastly, we examine the effect of the number of pairwise correlated features on the explanation accuracy of all linear models in our tabular datasets (Fig. 8). We only consider two features correlated if their Pearson pairwise correlation value is larger than 0.75. The number of correlated features increases



(a) Linear Regression (b) Logistic Regression (c) Naive Bayes

Fig. 9 Linear relationship between average explanation accuracy and the generalization of (**a**) Linear Regression (**b**) Logistic Regression (**c**) Naive Bayes models in synthetic datasets. Pearson correlation values, together with the p-values, are included in the legend. Note that the visualization of linear regression shows the mean squared error instead of the test accuracy



(a) Linear Regression (b) Logistic Regression (c) Naive Bayes

Fig. 10 Linear relationship between average explanation accuracy and model generalization of (**a**) Linear Regression (**b**) Logistic regression (**c**) Naive Bayes models for tabular datasets. Pearson Correlation values, together with the p-values, are included in the legend

the average explanation accuracy of LIME and SHAP for Logistic Regression and Naive Bayes models. Our experiments contradict the findings in the works of Molnar et al. (2022) and Gosiewska and Biecek (2019) in this case. In their studies, authors showed synthetic examples showing that correlated features in the dataset can contribute to low explanation accuracy in local explanations. One possible explanation for this incompatibility between our results and the aforementioned studies is that feature correlation can play a significant role when the feature vectors are correlated with the predicted output of explained models. As our explained models do not include interaction terms, these correlation values are negligible in our experiments.

5.2.3 Model generalization

Some studies (Molnar et al. 2022; Guidotti 2021) proposed that explanation accuracy increases positively with an increase in the predictive performance of models (the model generalization). The authors show synthetic datasets in which their proposed hypothesis hold. In this section, we study the linear relationship between the average explanation accuracy of datasets with the model test set accuracy for linear additive regression and classification models.

Similar to the previous section, we first examine the effect of model generalization in synthetic dataset settings. The synthetic dataset includes 40 features and four categorical variables, one-hot encoded into four bins. To investigate, we fit 20 different model variations of each explained linear additive model with different hyperparameters. In Fig. 9 , we can see that the average accuracy of all explanations of Linear Regression models decreases with an increase in model generalization.[11] All explanation of Logistic Regression and naive Bayes models have larger average accuracy for models that have larger test accuracy.

After that, we examine whether the same relationship holds in our tabular datasets. In Fig. 10, we see that the average accuracy of LIME and SHAP explanations has minimal changes with an increase in the generalization of Linear Regression models, while it decreases for LPI explanations of the same model. Moreover, the increase in model generalization negatively (positively) affects the accuracy of all local explanations of Logistic Regression (Naive Bayes) models.

The results from our synthetic and tabular datasets agree that the average accuracy of explanations of naive Bayes models increases with larger model test accuracy. In this case, our results are aligned with the findings in Molnar et al. (2022); Guidotti (2021) . However, we see opposite trends for all the explanations of Logistic Regression and LPI explanations of Linear Regression models. We can conclude that overall, the linear relationship between model generalization and explanation accuracy depends on the type of explained model and the explanation technique itself, similar to the effect of data as shown in Sect. 5.2.2.

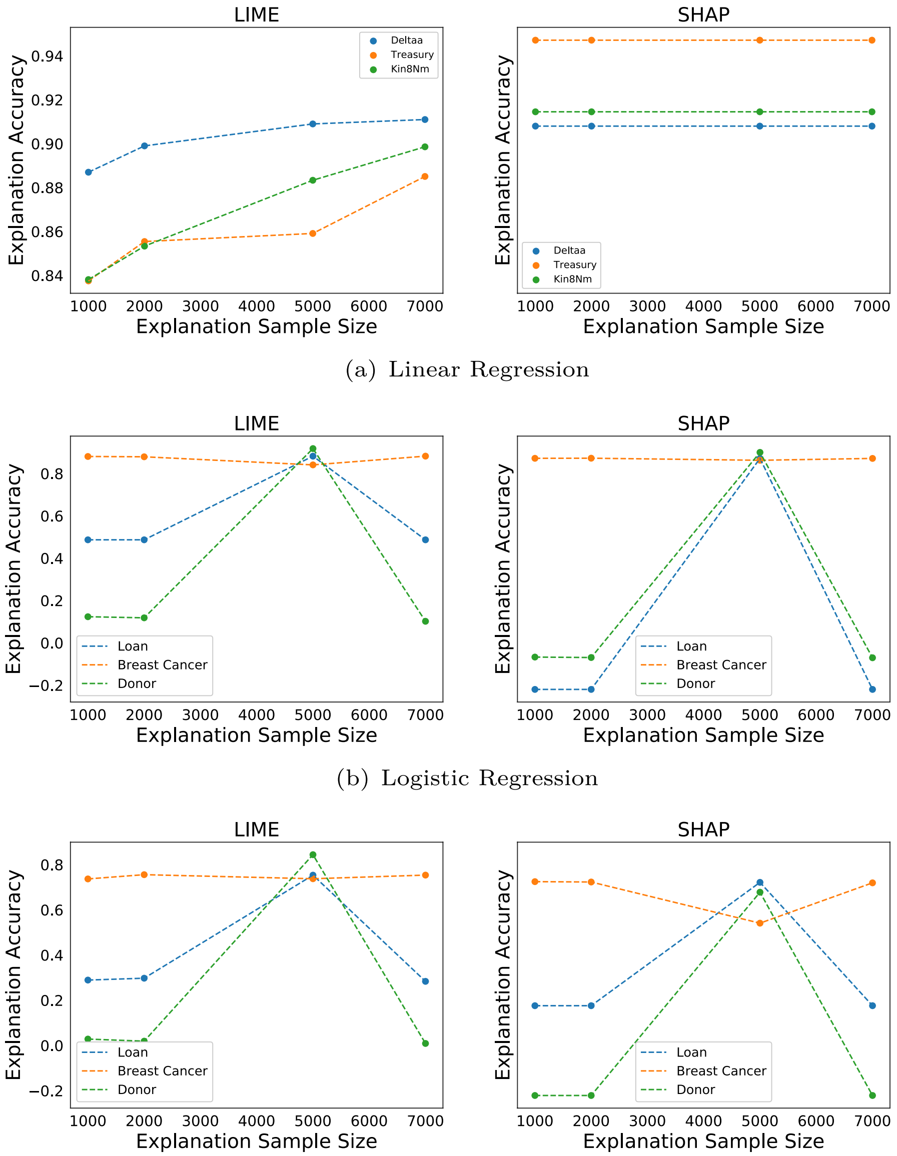
5.2.4 Explanation sample size

In explanations such as LIME and SHAP, the sample size is considered a hyperparameter that controls the number of samples that are generated in the locality of each explained instance[12]. One plausible assumption is that the larger the sample size, the higher the explanation accuracy. This is because sampling can provide more information about the local neighborhood of each instance to the explanation technique and, therefore, can increase the accuracy of the local explanation.

In this section, we study the relationship between the average explanation accuracy and the explanation sample size of LIME and SHAP techniques. For this experiment, we have included a subset of datasets based on how large is the size of their features. In Fig. 11, we can see the result of our experiments. Our results show our earlier hypothesis is correct only in LIME explanations of Delta A, Treasury, and Kin8NM datasets. Surprisingly, the average explanation accuracy of SHAP explanations of Linear regression models is constant across the selected regression datasets

[11] Note that the chart for linear regression includes the test error on the x-axis and not the accuracy, unlike the classification models.

[12] As mentioned earlier in Sects. 5.1.2 and 2.1, the sample size in LPI explanations is fixed and equal to the total number of test instances.



(a) Linear Regression

(b) Logistic Regression

(c) Naive Bayes

Fig. 11 The relationship between the explanation sample size of LIME and SHAP and the average explanation accuracy of LIME and SHAP across all datasets

for all sample sizes. When explaining Logistic Regression and Naive Bayes models, we can see that the average accuracy increases up to the sample size of 5000. However, the average accuracy of LIME and SHAP explanations follows a drastic decrease when the sample size is increased from 5000 to 7000 in Donors and Loan

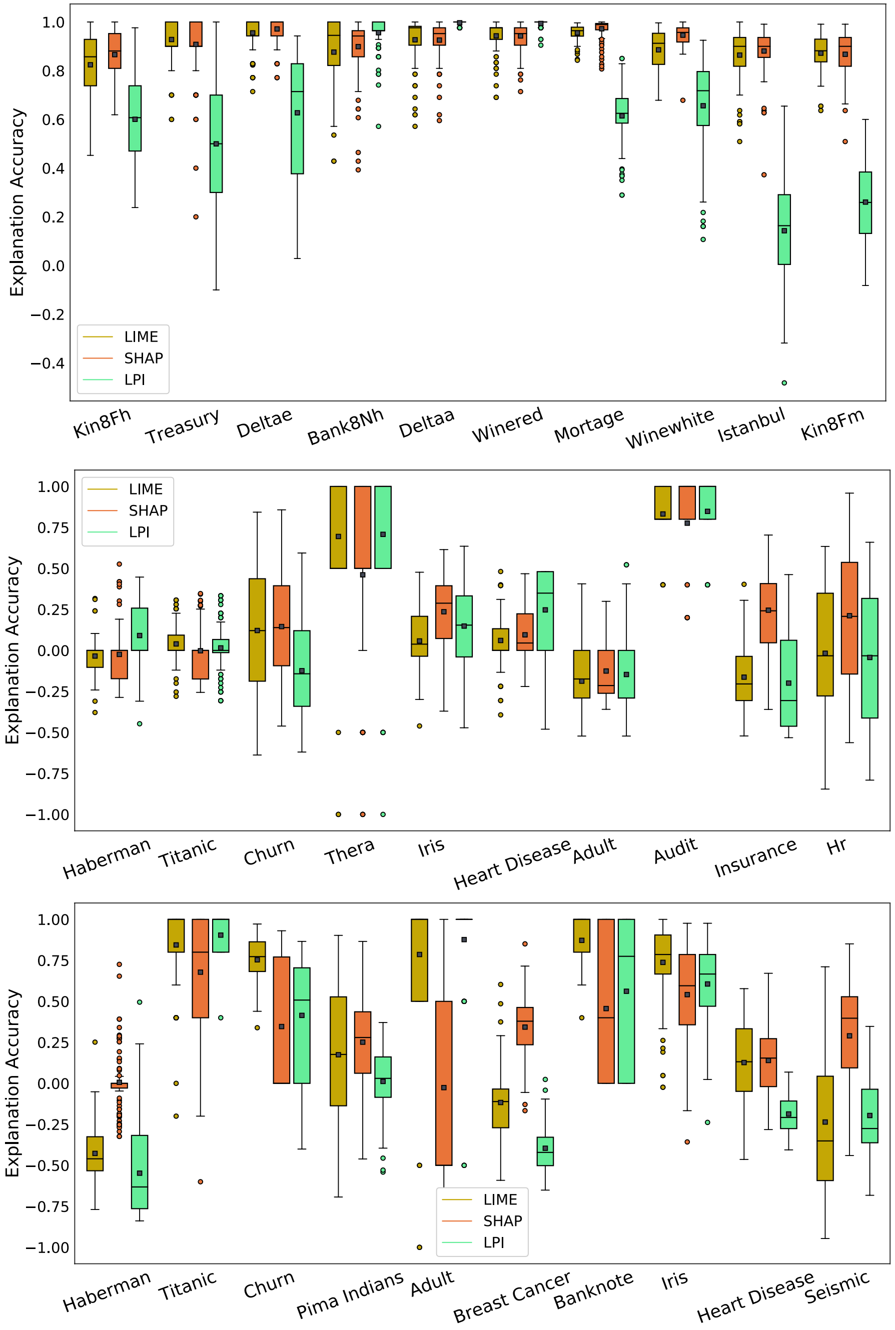


Fig. 12 Box-plots of explanation accuracy when the underlying model is Linear Regression (Top), Logistic Regression (Middle), and naive Bayes (Bottom). The dark rectangles are indicators of average values in each box plot

Table 5 The average of explanation accuracy across all data sets for different similarity metrics when explaining Linear Regression model

	LIME	SHAP	LPI
Spearman	0.904	**0.92**	0.507
Cosine	0.05	**0.075**	−0.342
Euclidean	0.464	**0.481**	0.47

Bold values indicate the explanation technique with the largest average accuracy

Table 6 The average explanation accuracy across all data sets based on the different similarity measures

	Logistic regression			Naive Bayes		
	LIME	SHAP	LPI	LIME	SHAP	LPI
Spearman	**0.258**	0.212	0.237	0.137	**0.228**	0.115
Cosine	0.008	0.036	**0.253**	0.012	0.031	**0.198**
Euclidean	0.495	**0.503**	0.484	0.228	**0.264**	0.256

Bold values indicate the explanation technique with the largest average accuracy

datasets. Note that the average local explanation accuracy can be affected significantly in the Donors and Loan datasets. This trend does not appear for LIME and SHAP explanations of the Breast Cancer dataset. One possible explanation can be that the Breast Cancer dataset only includes a few numerical features. We would like to reiterate that the choice of the sample size of 5000 in our study, as mentioned in Sect. 5.1.2 was to maximize the average explanation accuracy of these explanation techniques across all datasets and tasks.

According to our results, the relationship between explanation sample size and accuracy is more complicated than our former hypothesis. It is possible that our result can be explained by the findings in Laugel et al. (2018). The authors showed that increasing the explanation sample size can enlarge the neighborhood in the vicinity of an instance. Because of this, the surrogate model's decision boundary converges toward the global model's decision boundary. As a result, the local explanations converge towards global explanations, and therefore the *local* explanation accuracy decreases.

5.2.5 Variance in explanation accuracy

So far, we have focused on reporting the average explanation accuracy for each dataset. In our previous experiments, some explanations showed significantly low average explanation accuracy values. In certain cases, this can be explained by the large variance in their values of explanation accuracy. A large variance in explanation accuracy means the explanation technique can provide accurate explanations for a subset of instances and simultaneously provide inaccurate explanations for others. As mentioned in Sect. 4, since our evaluation technique can be performed at a single instance explanation level, we can measure the variance in average explanation accuracy within datasets.

Table 7 The test accuracy of Logistic Regression (LREG) and Naive Bayes (NB) models based on different preprocessing techniques used for each dataset

Dataset	Standard		Minmax		Robust	
	LREG	NB	LREG	NB	LREG	NB
Adult	0.85	0.56	0.85	0.55	0.85	0.58
Attrition	1	1	1	1	1	1
Audit	0.99	0.96	0.97	0.96	0.99	0.96
Banking	0.91	0.83	0.91	0.82	0.91	0.83
Banknote	0.98	0.85	0.98	0.85	0.98	0.85
Breast Cancer	0.97	0.92	0.97	0.92	0.97	0.92
Churn	0.81	0.8	0.81	0.8	0.81	0.8
Donor	1	1	1	1	1	1
Haberman	0.66	0.65	0.69	0.65	0.66	0.65
Hattrick	1	0.79	0.99	0.79	1	0.79
Heart Disease	0.83	0.8	0.83	0.8	0.83	0.8
Hr	0.78	0.38	0.78	0.37	0.78	0.38
Insurance	0.99	0.07	0.99	0.07	0.99	0.08
Iris	1	1	1	1	1	1
Loan	1	1	1	0.99	1	0.99
Pima Indians	0.8	0.77	0.78	0.77	0.8	0.77
Seismic	0.95	0.42	0.95	0.41	0.94	0.44
Spambase	0.92	0.81	0.89	0.81	0.89	0.81
Thera	1	1	1	0.99	1	0.99
Titanic	0.79	0.79	0.79	0.79	0.79	0.79

Note that the difference between the accuracy is negligible across all datasets

In Fig. 12, we show the top-10 datasets where the standard deviation in the explanation accuracy of all explanations is largest on average for each explained model. For example, LPI explanations show large standard deviations for Linear Regression models in datasets such as Treasury, Delta E, and Istanbul. Similar trends can be seen for the Logistic Regression explanations of Thera, Churn, and HR datasets. The standard deviation in explanation accuracy can be so significant for LIME explanations of the Adult dataset that the explanations range from the maximum to minimum accuracy in this dataset. Overall, the standard deviation in Naive Bayes explanations can be larger compared to the Logistic Regression explanations. Comparing the result in Fig. 12 with 1 can also show that large standard deviation in explanations of Naive Bayes is common among datasets in which the model has achieved a low generalization accuracy.

5.2.6 Choice of similarity measure

As discussed in Sect. 4.5, the wrong choice of similarity can draw misleading results. In this section, we evaluate to what extent the choice of similarity measure can affect the choice of the most accurate explanations across all discussed

Table 8 The average similarity across all datasets for each preprocessing technique

		Similarity	Logistic regression			Naive Bayes		
			LIME	SHAP	LPI	LIME	SHAP	LPI
Preprocessing	Standard	Spearman	**0.258**	0.212	0.237	0.137	**0.228**	0.115
		Cosine	0.008	0.036	**0.253**	0.012	0.031	**0.198**
		Norm	0.495	**0.503**	0.484	0.228	0.264	**0.256**
	Minmax	Spearman	0.165	**0.239**	0.067	−0.035	**0.188**	0.09
		Cosine	0.117	**0.119**	0.059	**0.202**	0.162	0.199
		Norm	**0.512**	0.505	0.489	0.231	**0.265**	0.254
	Robust	Spearman	0.192	**0.298**	0.176	0.063	**0.223**	0.085
		Cosine	0.057	0.072	**0.196**	0.011	0.031	**0.197**
		Norm	**0.507**	**0.507**	0.5	0.232	**0.264**	0.258

Bold values indicate the explanation technique with the largest average accuracy

models. Table 5 shows the average explanation accuracy of all explanations of Linear Regression across all datasets. Although SHAP outperforms other explanations for all similarity measures, we see that all explanations provide very similar average accuracy when using Euclidean similarity. For example, LPI can be preferred over LIME in case Euclidean similarity is the choice of similarity metric.

In Table 6, we can see that the choice of the most accurate explanation technique can largely be affected depending on the similarity metric used for classification models. This is why we have emphasized that awareness of the similarity metric used for evaluating local explanations is essential. Even though the choice of the correct similarity measure is highly dependent on the application scenario, in the context of tabular datasets, we argue that rank-based measures such as Spearman's rank correlation are the most appropriate, as proposed in Fong and Vedaldi (2017) and discussed earlier in Sect. 4.5.

5.2.7 Pre-processing effect

Since MIAS importance scores largely depend on the input feature values, we investigate the effect of the pre-processing techniques on our results of the explanation accuracy. We have realized that the effect of pre-processing on average explanation accuracy is significant for the Logistic Regression and Naive Bayes explanations even when their model generalization shows minimal change after each preprocessing technique is used on the dataset. For this experiment, we should highlight that we perform the pre-processing before training the explained model and use the preprocessed data when obtaining the local explanation. Table 7 shows the change in the test accuracy of classification models based on each pre-processing technique used. Note that pre-processing has little to no effect on the test accuracy of the two classification models.

We compare the average explanation accuracy of all local explanation techniques using these pre-processing techniques and with different similarity measures across

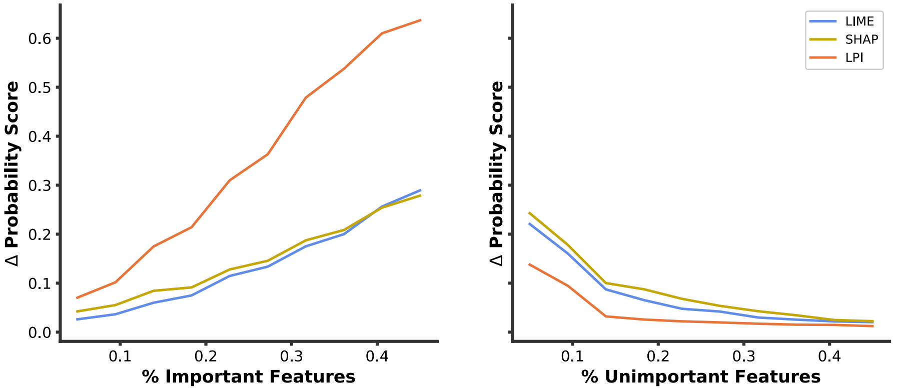


Fig. 13 Robustness of local explanations of logistic regression model for deletion (left) and Preservation (right) for the Breast Cancer Dataset

all datasets. The first three columns in Table 8 show the average accuracy across all datasets when explaining the Logistic regression model. SHAP provides the largest average explanation accuracy values for explanations of both models except for when min-max processing is used for Logistic Regression.

As mentioned earlier, in the case of tabular datasets, the most optimal choice of similarity metric is using Spearman's rank correlation when comparing additive and non-additive explanations. However, we can see that the combination of similarity metric and preprocessing techniques can significantly affect the choice of the most accurate explanations. For example, using Euclidean similarity can lead to choosing LIME for Logistic Regression explanations when min-max and robust preprocessing are used.

5.2.8 Explanation robustness

As we mentioned earlier in Sect. 1, we have evaluated the local explanations by measuring their similarity to our proposed MIAS scores directly. As we said earlier in Sect. 2.2.1 , most studies that evaluate local explanations have used the robustness measures. In this section, we provide experiments that evaluate the robustness of local explanations of linear additive models. By doing so, we aim to investigate whether average robustness measures are in agreement with the average explanation accuracy (Tables 3 and 4).

As we said earlier, the most popular measures of the robustness of local explanations are based on Deletion and Preservation (Hsieh et al. 2020; Montavon et al. 2018) . In these measures, we progressively nullify the top-K percent important (unimportant) features in the explained instance based on their importance in its local explanations. After that, we use the explained model as an oracle and obtain the change in the predicted probability score of the explained model on the new instance. As mentioned in Sect. 2.2.1, relatively large (small) values for Deletion (Preservation) measures indicate that the explanation are robust. Figure 13 shows

Table 9 Average robustness across all datasets based on the AUC measure for linear regression model

Model →	Preservation			Deletion		
Dataset	LIME	SHAP	LPI	LIME	SHAP	LPI
Anacalt	8.87	**8.13**	8.47	**46.64**	43.87	45.75
Bank8Fh	3.88	**3.48**	3.49	12.52	12.52	**12.57**
Bank8Fm	4.71	4.32	**4.29**	16.89	17.21	**17.23**
Bank8Nh	2.86	2.43	**2.36**	4.55	**5.16**	4.92
Bank8Nm	3.25	**2.6**	2.61	5.67	5.7	**5.73**
Delta A	**0.02**	**0.02**	**0.02**	0.02	0.02	**0.03**
Delta E	**0.11**	**0.11**	**0.11**	0.22	**0.23**	0.22
Istanbul	**0.64**	0.69	0.69	2.09	**2.24**	**2.24**
Kin8Fh	6.06	**5.38**	**5.38**	8.65	**8.63**	**8.63**
Kin8Fm	5.74	**5.22**	**5.22**	7.06	**8.09**	**8.09**
Kin8Nh	10.34	**10.29**	**10.29**	14.96	**15.27**	**15.27**
Kin8Nm	12.44	**9.34**	9.34	18.86	**21.2**	**21.2**
Mortage	248.31	**197.31**	205.49	**593.78**	560.04	532.67
Puma8Fh	150.94	**146.1**	**146.1**	613.8	**617.85**	**617.85**
Puma8Fm	100.22	99.58	**99.22**	463.35	**463.33**	463.7
Quakes	0.94	**0.92**	1.01	1.11	**1.17**	1.06
Treasury	372.65	**298.52**	316.97	688.04	**796.53**	746.1
Winered	19.63	**18.47**	31.97	76.16	**77.34**	56.43
Winewhite	33.52	**25.51**	46.85	100.87	**104.93**	88.83
Wizmir	523.29	**508.23**	508.42	2213.17	**2219.55**	2219.26
Average	75.38	**67.33**	70.42	244.41	**249.04**	243.39
Standard Deviation	140.62	128.12	129.77	506.96	512.38	509.51

Smaller values for preservation and larger values for deletion indicate larger explanation robustness

Bold values indicate the most robust explanation technique in each dataset

a visualization of the Breast Cancer dataset's Deletion and Preservation robustness measures for the Logistic Regression model averaged for all instances. In this case, LPI has the most robust explanation for the Logistic Regression based on both measures.

For calculating an overall measure of the robustness without relying on visualizations, Hsieh et al. (2020) proposed to calculate the AUC of figures similar to Fig. 13 . The formula they proposed is as follows: $AUC = \sum_{i=1}^{n}(y_i + y_{i-1})/2 * (x_i - x_{i-1})$. Based on this, robust explanations have the largest (smallest) AUC values concerning Deletion (Preservation) measures.

In Table 9, we can see that SHAP provides the most robust explanations on average across all datasets for Linear Regression explanations based on this AUC measure. In Tables 10 and 11 , we can see that LPI has the largest average robust explanations for Logistic Regression models and the Preservation measure for the Naive Bayes model. On the other hand, SHAP has the largest average robust explanations across all datasets for the Naive Bayes models.

Table 10 Average deletion robustness across all datasets based on the AUC measure

Model →	Logistic regression			Naïve Bayes		
Dataset	LIME	SHAP	LPI	LIME	SHAP	LPI
Adult	29.74	38.6	**41.8**	49.72	**54.22**	36.09
Attrition	67.04	66.76	**67.15**	**67.5**	**67.5**	**67.5**
Audit	34.63	33.4	**48.94**	106.88	**123.75**	**123.75**
Banking	9.2	9.94	**11.79**	27.32	**28.45**	20.67
Banknote	72.04	72.51	**77.63**	25.74	26.49	**28.04**
Breast Cancer	23.76	13.73	**73.34**	4.63	**7.46**	4.38
Churn	20.59	**21.72**	20.28	27.11	**27.88**	20.69
Donor	67.27	**67.29**	67.28	**67.5**	**67.5**	**67.5**
Haberman	**7.6**	5.71	7.54	**2.97**	2.29	**2.97**
Hattrick	63.13	61.83	**68.75**	27.38	**39.85**	19.9
Heart Disease	28.53	**45.18**	34.19	54.61	**68.31**	59.57
Hr	**39.41**	31.51	27.84	**67.5**	**67.5**	**67.5**
Insurance	1.22	**1.54**	0.96	**0**	**0**	**0**
Iris	17.17	16.57	**18.63**	**45**	**45**	**45**
Loan	44.89	44.86	**44.91**	**45**	**45**	**45**
Pima Indians	**43.35**	42.55	42.47	47.68	**53.45**	47.21
Seismic	8.9	**9.17**	8.42	86.08	**90.41**	65.1
Spambase	50.55	30.82	**55.92**	48.38	145.13	**146.25**
Thera	44.88	44.84	**44.9**	**45**	**45**	**45**
Titanic	56.33	**57.28**	50.44	52.03	**64.83**	47.05
Average	36.51	35.79	**40.66**	44.9	**53.5**	47.96
Standard deviation	21.36	21.57	23.16	26.41	35.81	36.09

Larger values indicate higher robustness for this measure

Bold values indicate the most robust explanation technique in each dataset

Not only we see a different trend in the average explanation robustness across all datasets, we see many differences between explanation accuracy and explanation robustness across all of our models (See Tables 4 and 3). For example, we can see that the average robustness values of explanations is equal for HR, Insurance, Iris, and Loan when explaining naive Bayes models. However, we can see that the average explanation accuracy differs for the explanations of naive Bayes models. These differences are also visible in the explanation robustness of Linear Regression models. LIME explanations provided the largest average accuracy in the Delta A dataset. In contrast, SHAP and LPI are the most robust explanations concerning Preservation and Deletion. Based on our result, we can conclude that the most robust explanations of linear additive models based on Deletion and Preservation do not necessarily have the largest average accuracy and vice versa.

Table 11 Average robustness based on the preservation measure across all datasets

Model →	Logistic regression			Naïve Bayes		
Dataset	LIME	SHAP	LPI	LIME	SHAP	LPI
Adult	27.79	**17.56**	24	30.91	**22.12**	27.22
Attrition	**0.25**	0.73	0.31	**0**	**0**	**0**
Audit	19.69	38.13	**19.23**	**123.75**	**123.75**	**123.75**
Banking	**4.39**	6.09	4.72	6.82	**3.84**	10.29
Banknote	17.09	**15.45**	16.17	16.63	**15.8**	18.12
Breast cancer	11.2	26.94	**5.04**	**0**	0.05	1.69
Churn	**11.6**	12.96	13.16	14.5	**13.43**	21.61
Donor	0.05	0.06	**0.03**	**0**	**0**	**0**
Haberman	**9.07**	11.03	9.11	**4.03**	4.69	**4.03**
Hattrick	**0.36**	11.52	1.16	4.83	**0.77**	15.05
Heart disease	25.63	**16.76**	21.91	41.7	41.06	**40.88**
Hr	**19.11**	30.87	28.6	**0.01**	**0.01**	**0.01**
Insurance	1.18	**0.66**	1.63	**0**	**0**	**0**
Iris	31.72	31.95	**31.04**	**11.26**	**11.26**	**11.26**
Loan	**0.17**	0.23	**0.17**	**0**	**0**	**0**
Pima Indians	12.83	**12.47**	13.68	23.76	**19.29**	24.58
Seismic	3.63	**2.44**	4.76	6.59	**4.79**	26.93
Spambase	28.07	48.64	**20.05**	146.25	146.25	**14.63**
Thera	**0.17**	0.24	0.18	**0**	**0**	**0**
Titanic	**19.51**	23.87	25.73	31.68	**21.15**	40.51
Average	12.18	15.43	**12.03**	23.14	21.41	**19.03**
Standard deviation	10.55	13.74	10.38	39.37	39.46	27.33

Smaller values indicate higher robustness with respect to this measure

Bold values indicate the most robust explanation technique in each dataset

6 Discussion

We have identified some limitations of our study. Firstly, our conclusions are limited to explaining three linear additive models: Linear and Logistic Regression and Naive Bayes. Therefore, we cannot generalize these results to the accuracy of local explanations of more complex black box models and other additive models. Secondly, our proposed functionally-grounded evaluation of local explanations cannot replace the human-grounded evaluation procedures. As we said earlier, the functionally-grounded evaluation methods, such as the one presented here, can only be seen as means to make new candidate techniques more efficient; they allow for rejecting some candidate techniques early on. Suppose explanation techniques pass our sanity checks, such as LIME and SHAP explanation of Linear Regression in our regression datasets. In that case, they will likely still need to be qualitatively evaluated in a user-centered context later.

Since our study was focused on evaluating local explanations of linear additive models, the question that might occur to the reader is whether the inaccuracy of the explanations for linear additive models can tell us anything about the accuracy of local explanations of more complex models. Even though our conclusion about the performance on simpler models does not necessarily transfer to good performance on complex tasks, we argue that *if the explanation technique is not accurate for simpler models, then it is very unlikely that it will have high accuracy on complex models*.

7 Concluding remarks

Our study proposed a sanity check that examined whether local additive model-agnostic explanations can provide accurate explanations for linear additive models. The evaluation was based on extracting Model-Intrinsic Additive Scores (MIAS) from linear additive models such as linear and logistic regression and naive Bayes models. We then used the similarity of our proposed scores with local explanations using Spearman's rank correlation to measure explanation accuracy.

It can be intuitive to assume that local additive explanations could provide high-accuracy explanations and pass our sanity check. However, we showed that this is not always the case. While LIME and SHAP explanations of Linear Regression models do pass our sanity check, they failed to provide accurate local explanations across numerous datasets for Logistic Regression and Naive Bayes models. We can conclude that these local explanations cannot be trusted in high stake decision-making cases in their current state for classification tasks.

One possible explanation for the fact that these explanations fail our sanity check for Logistic Regression and Naive Bayes models is that both LIME and SHAP explain the predicted probability scores of a designated class using linear regression surrogates. For this, we suggest investigating the use of classification surrogates in these explanations when explaining classification models. Another explanation can be that Logistic Regression and Naive Bayes models have decision boundaries that are more complex than Linear Regression models. For this, we suggest future studies to investigate the relationship between the model complexity and the accuracy of local explanations. We hope that future studies can use our evaluation method and sanity check in this endeavour.

In our study, we also examined whether additive explanations such as LIME and SHAP are more accurate than non-additive LPI explanations when explaining linear additive models. Our empirical investigation showed that while this is true for Linear Regression models, LPI explanations have larger average accuracy in a subset of our studied datasets for Logistic Regression and Naive Bayes models. Therefore, we can conclude that in some cases, the additivity of explanations is not necessarily an advantage for explaining linear additive models.

We provided an extensive analysis of the factors that may affect explanation accuracy. Our results show that the accuracy of the explanation techniques can depend on the number of numerical and categorical features, pairwise feature correlation, model generalization, similarity metric, pre-processing techniques, and explanation sample size. We showed that the effect of these factors on explanation accuracy is highly dependent on the type of model we explain and the explanation technique. Using this knowledge, we can set control mechanisms for the factors affecting each explanation when evaluating local explanations.

In their current stage, LIME and SHAP have no criteria for when they should not provide explanations. Based on the significant standard deviation in the explanation accuracy of linear additive models, we argue that these explanations need internal mechanisms to stop explaining when facing uncertainty for achieving high accuracy on a dataset level.

Our evaluation method requires that the prediction function of linear additive classification models could be transformed into a linear additive model. In principle, just like our proposed log odds trick (In Sects. 4.3 and 4.4), by transforming the prediction function of any machine learning into linear additive models, our evaluation method can be used to calculate the accuracy of local explanations directly. One important direction for future research is to extend the proposed evaluation framework to other model classes, e.g., tree models, and explanation types, e.g., rules, as produced by Anchors (Ribeiro et al. 2018). One major challenge here is to derive the model-intrinsic feature importance scores in cases where intrinsic additive structures are not as easily derivable as they are for logistic regression and naive Bayes.

Another important direction for future studies is to evaluate the accuracy of local explanations of linear additive models for other modalities of data, such as text and images. In those cases, the challenge is finding datasets where linear additive models provide acceptable accuracy for obtaining their local explanations. Since our evaluation method of explanations is designed for local additive explanations, we do not recommend its use for evaluating local explanations that are not instrinsically additive.

Appendix

Logistic regression example

To make our idea more tangible, we show an example of extracting MIAS scores for Logistic Regression. We train a logistic regression model with L2 regularization on a 2-dimensional discrete *logical AND* function that returns one if both inputs are one and zero otherwise. The parameters of the logistic regression model are $w^1 = 0.422$ and $w^2 = 0.422$ with the intercept value $w_0 = 0.69$. These parameters show that the model correctly learned that both features are equally important on a global level. For x_0 with $x_0^1 = 1$ and $x_0^2 = 0$ the model predicts $P(y_0 = 1 \mid x_0) = 0.75$. Based on this, we can derive the log odds ratio for x_n as:

$$\log\left(\frac{0.75}{0.25}\right) = 0.69 + 0.422 \times 1 + 0.422 \times 0$$

We can see that, whereas the first and second feature are equally important globally, the only feature that contributes to the log odds prediction is the first feature. The resulting MIAS scores are $\lambda_0^1 = 0.422$ and $\lambda_0^2 = 0$, different from the global explanation used by studies such as Agarwal et al. (2022). A similar example of the naive Bayes model can be found in the Appendix.

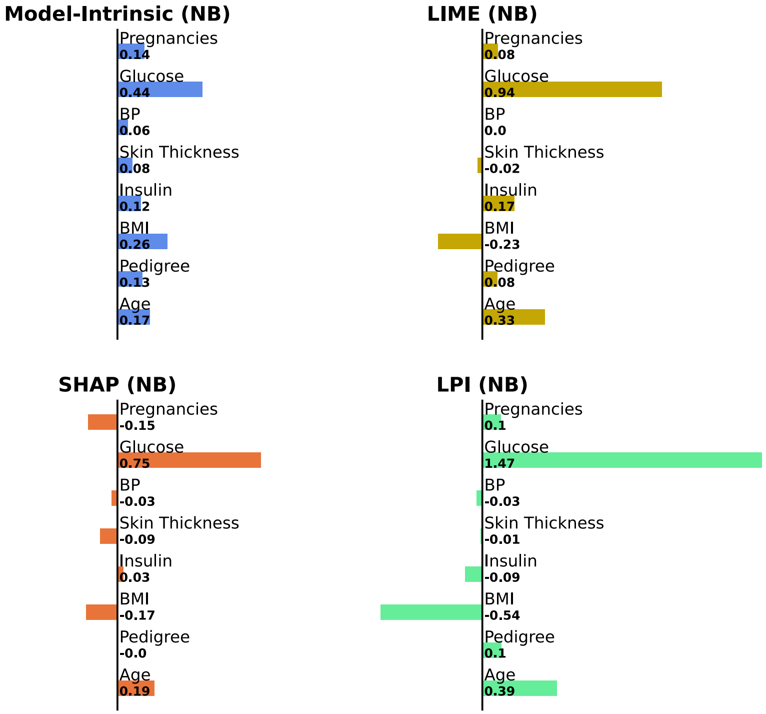


Fig. 14 The explanations of LIME, SHAP and LPI explanations for a single instance from the Pima Indians data set along with MIAS scores when explaining a Naive Bayes prediction

Naïve Bayes example

In this section, we show an example of how LOR scores are extracted for a Naive Bayes model. Let us train a Gaussian Nav̈e Bayes model on the following data and label matrix:

$$X = \begin{pmatrix} -1 & -1 \\ -2 & -1 \\ -3 & -2 \\ 1 & 1 \\ 2 & 1 \\ 3 & 2 \end{pmatrix}, Y = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

The parameters of the Gaussian distribution for feature 1 and 2 for class 0 are $\mathcal{N}(-2.0, 0.66)$ and $\mathcal{N}(-1.33, 0.22)$. Similarly, parameters of the Gaussian distribution for features 1 and 2 for class 1 are: $\mathcal{N}(2.0, 0.66), \mathcal{N}(1.33, 0.22)$. For x_n with $x_n^1 = -2$ and $x_n^2 = -1$, the model predicts the class to be 1 with probability 1. Let $c = 0$, therefore,

Table 12 The model accuracy for our synthetic experiments

Experiment	Model	Trial numbers								
		1	2	3	4	5	6	7	8	9
Numerical	Logistic regression	0.9	0.875	0.85	0.825	0.85	0.8	0.825	0.775	0.925
	Naive Bayes	0.875	0.875	0.8	0.875	0.85	0.925	0.825	0.875	0.975
	Linear regression	0.9	0.875	0.85	0.825	0.85	0.8	0.825	0.775	0.925
Categorical	Logistic regression	0.775	0.85	0.875	0.85	0.65	0.7	0.675	–	–
	Naive Bayes	0.875	0.725	0.7	0.625	0.7	0.65	0.725	–	–
	Linear regression	3.0e−02	6.2e+03	1.6e+05	3.7e+04	1.9e+04	2.8e+04	2.6e+04	–	–

The results for linear regression model is based on mean square error

Table 13 The F-statistic and p-values from the comparison of explanations based on their explanation accuracy using different similarity measures

	Models	
	F-Stastic	p-Value
Linear regression	6.87, 0.032	
Logistic regression	0.363	0.83
Naive Bayes	0.363	0.833

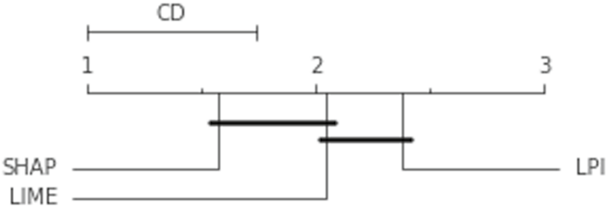


Fig. 15 Post-hoc Nemenyi test for the explanation accuracy of Linear Regression models. The line connects the samples with no significant differences

$$\mathcal{N}(x_n^0||\mu_c^0, \sigma_c^0) = 0.488$$
$$\mathcal{N}(x_n^1||\mu_c^1, \sigma_c^1) = 3.002e^{-6}$$
$$\mathcal{N}(x_n^0||\mu_{\neg c}^0, \sigma_{\neg c}^0) = 0.65$$
$$\mathcal{N}(x_n^1||\mu_{\neg c}^1, \sigma_{\neg c}^1) = 4.04e^{-6}$$

based on this,

$$\begin{aligned} \log\left(\frac{1}{3.7e^{-11}}\right) =& \log\frac{0.488}{3.002e^{-6}} \\ &+ \log\frac{0.65}{4.04e^{-6}} \\ 23.99 =& 11.99 + 11.99 \end{aligned}$$

where $const = log(1)$. While the first feature has an average of −2 for class 0 in the global Gaussian distribution parameters, the contribution of this feature to the LOR of Nav̈e Bayes model for x_n is largely positive, i.e. 11.99. From these two examples, we can see that beside providing true local importance scores, our proposed method can help to quantify the differences between the global and local importance scores in Naive Bayes. Figure 14 shows another example where the extracted LOR scores are compared against different explanations for a test instance in Pima Indians dataset.

The data effect

We studied the effect of numerical and categorical features in Sect. 5.2.2. The test accuracy of the models trained for those trials are reported in Table 12.

Statistical significance

In this section, we investigate whether the performance of explanation techniques is significantly different from one another based on statistical testing. To investigate this question, we compare the average accuracy of explanation techniques across both classification and regression datasets using the Friedman test as proposed in Demšar (2006). The null hypothesis of the Friedman test is that there are no significant differences between the samples of the average explanation accuracy values. In Table 13, we can see the result of the F-Statistic and the corresponding *p*-values. Based on the results, we can only reject the null hypothesis in the case of the explanations of Linear Regression models. For this case, and in order to investigate the pairwise differences, we continue with the post-hoc Nemenyi test. Figure 15 shows significant pairwise differences between the explanation accuracy of SHAP and LPI.

Datasets

Descriptions and access links to all datasets used in this study are available in Tables 14 and 15.

Table 14 Classification datasets

Dataset	Source
Adult	UCI
Attrition	Kaggle
Audit	Kaggle
Banking	Kaggle
Banknote	UCI
Breast cancer	UCI
Churn	Kaggle
Donors	Kaggle
Pima Indians	Kaggle
Haberman	UCI
Hattrick	Kaggle
Heart disease	UCI
HR	Kaggle
Insurance	Kaggle
Iris	UCI
Loan	Kaggle
Seismic	UCI
Spambase	UCI
Thera	Kaggle
Titanic	Kaggle

Table 15 Regression datasets

Dataset	Source
Anacalt	KEEL
Bank8Fh	Delve
Bank8Fm	Delve
Bank8Nh	Delve
Bank8Nm	Delve
Delta A	KEEL
Delta E	KEEL
Istanbul	UCI
Kin8Fm	Delve
Kin8Nh	Delve
Kin8Fh	Delve
Kin8Nm	Delve
Mortgage	KEEL
Puma8Fh	Delve
Puma8Fm	Delve
Quakes	KEEL
Treasury	KEEL
Wine red	UCI
Wine white	UCI
Wizmir	KEEL

Funding Open access funding provided by Royal Institute of Technology.

References

Aas K, Jullum M, Løland A (2021) Explaining individual predictions when features are dependent: more accurate approximations to shapley values. Artif Intell 298:103502

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. arXiv preprint arXiv:1810.03292

Agarwal C, Krishna S, Saxena E, Pawelczyk M, Johnson N, Puri I, Zitnik M, Lakkaraju H (2022) Openxai: towards a transparent evaluation of model explanations. Adva Neur Inform Process Syst 35:15784–15799

Alvarez Melis D, Jaakkola T (2018) Towards robust interpretability with self-explaining neural networks. Advances in neural information processing systems 31

Alvarez-Melis D, Jaakkola TS (2018) On the robustness of interpretability methods. ICML Workshop on human interpretability in machine learning

Amparore E, Perotti A, Bajardi P (2021) To trust or not to trust an explanation: using leaf to evaluate local linear XAI methods. PeerJ Comput Sci 7:479

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Casalicchio G, Molnar C, Bischl B (2018) Visualizing the feature importance for black box models. In: Joint European conference on machine learning and knowledge discovery in databases, pp. 655–670. Springer

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608

Faber L, Moghaddam AK, Wattenhofer R (2021) When comparing to ground truth is wrong: On evaluating gnn explanation methods. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp. 332–341

Fong RC, Vedaldi A (2017) Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE international conference on computer vision, pp. 3429–3437

Freitas AA (2014) Comprehensible classification models: a position paper. ACM SIGKDD Explorat Newsl 15(1):1–10

Ghorbani A, Abid A, Zou J (2019) Interpretation of neural networks is fragile. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33, pp. 3681–3688

Gosiewska A, Biecek P (2019) Do not trust additive explanations. arXiv preprint arXiv:1903.11420

Guidotti R (2021) Evaluating local explanation methods on ground truth. Artif Intell 291:103428

Hakkoum H, Abnane I, Idri A (2022) Interpretability in the medical field: a systematic mapping and review study. Appl Soft Comput 117:108391

Hooker S, Erhan D, Kindermans P-J, Kim B (2019) A benchmark for interpretability methods in deep neural networks. Advances in Neural Information Processing Systems 32 (NeurIPS)

Hsieh C-Y, Yeh C-K, Liu X, Ravikumar P, Kim S, Kumar S, Hsieh C-J (2020) Evaluations and methods for explanation through robustness analysis. arXiv preprint arXiv:2006.00442

Kramer O, Kramer O (2016) Scikit-learn. Machine learning for evolution strategies, 45–53

Lakkaraju H, Arsov N, Bastani O (2020) Robust and stable black box explanations. In: International conference on machine learning, pp. 5628–5638. PMLR

Laugel T, Renard X, Lesot M-J, Marsala C, Detyniecki M (2018) Defining locality for surrogates in post-hoc interpretablity. arXiv preprint arXiv:1806.07498

Liu Y, Khandagale S, White C, Neiswanger W (2021) Synthetic benchmarks for scientific research in explainable machine learning. arXiv preprint arXiv:2106.12543

Liu M, Mroueh Y, Ross J, Zhang W, Cui X, Das P, Yang T (2019) Towards better understanding of adaptive gradient algorithms in generative adversarial nets. arXiv preprint arXiv:1912.11940

Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. Advances in neural information processing systems 30 (NeruIPS)

Molnar C, König G, Herbinger J, Freiesleben T, Dandl S, Scholbeck CA, Casalicchio G, Grosse-Wentrup M, Bischl B (2022) General pitfalls of model-agnostic interpretation methods for machine learning models. In: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, pp. 39–68. Springer

Montavon G, Samek W, Müller K-R (2018) Methods for interpreting and understanding deep neural networks. Digit Signal Process 73:1–15

Nguyen A-p, Martínez MR (2020) On quantitative aspects of model interpretability. arXiv preprint arXiv:2007.07584

Omeiza D, Speakman S, Cintas C, Weldermariam K (2019) Smooth grad-cam++: an enhanced inference level visualization technique for deep convolutional neural network models. arXiv preprint arXiv:1908.01224

Plumb G, Molitor D, Talwalkar AS (2018) Model agnostic supervised local explanations. Advances in neural information processing systems 31

Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H (2021) Manipulating and measuring model interpretability. In: Proceedings of the 2021 CHI conference on human factors in computing systems, pp. 1–52

Rahnama AHA, Boström H (2019) A study of data and label shift in the lime framework. Neurip 2019 Workshop on human-centric machine learning

Ribeiro MT, Singh S, Guestrin C (2016) "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, pp. 1135–1144

Ribeiro MT, Singh S, Guestrin C (2016) Model-agnostic interpretability of machine learning. ICML Workshop on human interpretability in machine

Ribeiro MT, Singh S, Guestrin C (2018) Anchors: high-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence, vol. 32

Ross SM (2017) Introductory statistics. Academic Press, Cambridge

Rudin C (2018) Please stop explaining black box models for high stakes decisions. Stat, 1050:26

Samek W, Binder A, Montavon G, Lapuschkin S, Müller K-R (2016) Evaluating the visualization of what a deep neural network has learned. IEEE Trans Neural Netw Learn Syst 28(11):2660–2673

Shrikumar A, Greenside P, Shcherbina A, Kundaje A (2016) Not just a black box: learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713

Sturmfels P, Lundberg S, Lee S-I (2020) Visualizing the impact of feature attribution baselines. Distill 5(1):22

van der Waa J, Nieuwburg E, Cremers A, Neerincx M (2021) Evaluating xai: a comparison of rule-based and example-based explanations. Artif Intell 291:103404

Wang C, Han B, Patel B, Rudin C (2022) In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. J Quantit Criminol 39(2):519–581

Yang M, Kim B (2019) Benchmarking attribution methods with relative feature importance. Neurip 2019 workshop on human-centric machine learning

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision, pp. 818–833. Springer

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Amir Hossein Akhavan Rahnama[1] · Judith Bütepage[1] · Pierre Geurts[2] · Henrik Boström[1]

✉ Amir Hossein Akhavan Rahnama
amiakh@kth.se

Judith Bütepage
butepage@kth.se

Pierre Geurts
p.geurts@uliege.be

Henrik Boström
bostromh@kth.se

[1] Department of Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

[2] Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium