LEAFAGE

Example-based and Feature importance-based Explanations for Black-box ML models

Adhikari, Ajaya; Tax, David M.J.; Satta, Riccardo; Faeth, Matthias

DOI
10.1109/FUZZ-IEEE.2019.8858846

Publication date
2019

Document Version
Final published version

Published in
2019 IEEE International Conference on Fuzzy Systems, FUZZ 2019

LEAFAGE: Example-based and Feature importance-based Explanations for Black-box ML models

Ajaya Adhikari
Data Science Department
TNO
The Hague, The Netherlands
ajaya.adhikari@tno.nl

David M. J. Tax
EEMCS
Delft University of Technology
Delft, The Netherlands
D.M.J.Tax@tudelft.nl

Riccardo Satta
Data Science Department
TNO
The Hague, The Netherlands
riccardo.satta@tno.nl

Matthias Faeth
Data Science Department
TNO
The Hague, The Netherlands
matthias.faeth@tno.nl

***Abstract*—Explainable Artificial Intelligence (XAI) is an emergent research field which tries to cope with the lack of transparency of AI systems, by providing human understandable explanations for the underlying Machine Learning models. This work presents a new explanation extraction method called LEAFAGE. Explanations are provided both in terms of feature importance and of similar classification examples. The latter is a well known strategy for problem solving and justification in social science. LEAFAGE leverages on the fact that the reasoning behind a single decision/prediction for a single data point is generally simpler to understand than the complete model; it produces explanations by generating simpler yet locally accurate approximations of the original model. LEAFAGE performs overall better than the current state of the art in terms of fidelity of the model approximation, in particular when Machine Learning models with non-linear decision boundaries are analysed. LEAFAGE was also tested in terms of usefulness for the user, an aspect still largely overlooked in the scientific literature. Results show interesting and partly counter-intuitive findings, such as the fact that providing no explanation is sometimes better than providing certain kinds of explanation.**

***Index Terms*—eXplainable AI, example-based reasoning, empirical study**

I. Introduction

In the context of Artificial Intelligence, Machine Learning (ML) is a rapidly growing field. There has been a surge of high-performance models for classification and prediction. Still, the application of these models in high-risk domains is more stagnant due to lack of transparency and trust: there is a disconnect between the black-box character of these models and the needs of the users. Explainable Artificial Intelligence (XAI) has recently emerged to provide solutions to this issue by attempting to create understandable explanations for the reasoning of a black-box model.

Example-Based Reasoning (EBR), i.e., motivating a decision by providing examples of similar situations, is widely recognized as an effective way to provide explanations [1], as it bears a close resemblance to the way humans think. As a result, it is commonly used e.g. in the health-care sector for decision-support systems [2], [3] and in law for justifying arguments, positions and decisions [4]. However, the usage of EBR to explain black-box ML models (i.e., models whose inner mechanisms are either unknown by the user, or too complex to be practically comprehensible by a human), has been largely overlooked so far in the scientific literature. This is partly because of the difficulty of finding examples according to the inner reasoning of such a model. Notably, most of the scientific literature focuses instead on evaluating the relative importance of features (*feature importance-based* explanations, see e.g. LIME [5]).

In this paper, we propose a new method for providing both feature importance-based and EBR explanations of the local reasoning of black-box models. Here, *local* refers to the ability of tailoring the explanation to a single prediction taken by the ML model, as opposed to providing a global explanation of the whole model logic. We named the method *LEAFAGE* - Local Example and Feature importance-based model AGnostic Explanations. LEAFAGE approximates the local reasoning of the black-box model by a (transparent) linear model. As a byproduct, LEAFAGE is also able to provide the importance of each feature for a prediction.

We evaluate LEAFAGE both in terms of fidelity, and of usefulness to the user. *Fidelity* refers to whether the extracted explanation reflects the true reasoning of the underlying black-box ML model. The *usefulness* to the user is evaluated by conducting a user-study in terms of perceived aid in decision-making and objective transparency.

The remainder of this paper is structured as follows. In Chapter II, we provide background information and related work on XAI, and explore approaches to provide explanations that leverage on social research. Chapter III describes LEAFAGE, which is then evaluated in terms of fidelity and usefulness to the user in Chapters IV and V. Finally, Chapter VI draws conclusions and suggests future research directions.

II. Background

This Section surveys current literature on XAI for ML models, and on the user's perspective on an explanation.

978-1-5386-1728-1/19/$31.00

An explanation about a ML model can be of *global* or *local* scope. A *global* explanation clarifies the inner workings of the whole ML model, i.e., how the relationship between input and output spaces is modeled [6]. *Local* explanations look instead at the reasoning behind a decision/prediction over a single input data point (*test* sample), thus targeting a sub-region of the input space. As the complexity of the ML model grows, it becomes harder to generate an understandable global explanation. However, it is likely that the logic of the ML model in the neighbourhood of a single test sample will be much simpler, thus allowing to generate understandable *local* explanations.

Three main strategies for extracting human-understandable explanations from ML models can be found in the literature: *transparent-by-design*, *model-oriented* and *model-agnostic*. In the first strategy, the ML model is designed from the start to be globally transparent and possibly simple enough to be understandable by humans (e.g. a small decision tree). The latter two strategies deal instead with an existing model that has not been made transparent by design. In the model-oriented strategy, certain parts of the model are used to extract an explanation (e.g., see [7]). In case when the ML model is too complex, or when internal workings of the model are not accessible, a model-agnostic strategy is used. This strategy views the ML model as a *black-box*, and queries it using a set of instances from the input space in order to gain insights in the behaviour of the model.

The proposed method, LEAFAGE, falls into the latter category. One of the most recent methods on the same category is LIME (Local Interpretable Model-agnostic Explanations) [5], from which LEAFAGE borrows its main ideas. LIME provides a *local explanation* by linearly approximating the decision boundary of the ML model in the neighbourhood of the test sample. Figure 1 shows an example of how LIME works in a binary classification problem. The two classes are Red and Blue, respectively represented by '+' and full circles. The decision boundary of the ML model is between yellow/blue areas. The point marked with a bold '+' is the test sample $\mathbf{z}$. In order to generate an explanation as for why the model ML classified $\mathbf{z}$ as Red, other Red and Blue synthetic data points are sampled from the input space.

A linear model is learned on the synthetic data points; higher importance is given in correctly classifying the synthetic instances that are close to $\mathbf{z}$. In Figure 1 their size represents their proximity to $\mathbf{z}$. This proximity from an synthetic instance $\mathbf{x}$ to $\mathbf{z}$ is defined by an exponential kernel $\pi(\mathbf{z}, \mathbf{x}) = e^{-D(\mathbf{x},\mathbf{z})^2/\sigma^2}$. In [5], σ is fixed to $0.75 * \sqrt{dimension}$.

The parameter σ of the proximity kernel plays an important role: a fixed σ can lead to neighbourhoods that do not include a decision boundary, or that include a too big part of the decision boundary which cannot be approximated linearly. Laugel et al. [8] spotted this problem and suggested to sample instances close to the nearest decision boundary to $\mathbf{z}$ within a fixed hyper-sphere. However, the fixed hyper-sphere can also lead to too small or too big neighbourhoods.

XAI research has been criticized for overlooking the view-

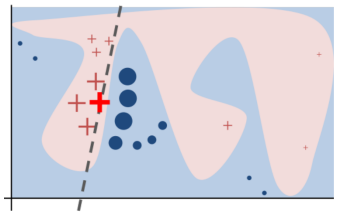


Figure 1: The dashed line approximates the blue/red decision boundary in the neighbourhood of the bold '+' point [5].

point of the end-user, i.e., if he/she is satisfied with provided explanations [9]. To address this, the present work focuses on providing explanations according to Example-Based Reasoning (EBR), a paradigm where explanations are related to previous experience [4], [10]. This type of reasoning lies very close to how humans think [1], [3].

EBR applications can be divided into two types: *problem-solving* and *decision-justification* [4]. In *problem-solving*, previous similar situations are used as aid to decide how to proceed with the current situation. In *decision-justification*, previous similar situations are leveraged to support or dismiss certain arguments and decisions. Worth noting, Common Law, which is used in most English-speaking countries, is based on the same principle (judicial decision are made on similar cases from the past [10]).

The ultimate goal of an ML explanation system is to provide valuable insights on an automated prediction/decision to the user. Such aspect can be evaluated by conducting user-studies. In recommendation systems, extensive research has been conducted in designing user-studies which evaluate explanations that clarify why a certain item is recommended, from the user's point of view [11]–[13]. Tintarev [11] defines seven goals for an explanation system, namely *transparency*, *scrutability*, *trust*, *effectiveness*, *efficiency*, *persuasiveness* and *satisfaction*. All of them can be evaluated subjectively by asking questions to the user [11]. However, while *trust*, *satisfaction* and *effectiveness* are subjective by nature, *transparency*, *efficiency* and *persuasiveness* can also be measured objectively. E.g., one can test whether users have understood the reasoning behind the recommendations [13], measure the interaction time [12] or check whether the user agrees to buy a recommended item.

III. LEAFAGE

This Section describes our proposed method, LEAFAGE (Local Example and Feature importance-based model AGnostic Explanation). It provides explanations in the form of examples drawn from the training set, that are similar to the test sample *according to the ML model logic*, and shows the importance of each feature for the prediction.

Let $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a black-box ML model that solves a binary classification problem with $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{c_1, c_2\}$, and $\mathbf{z} \in \mathcal{X}$ be the test sample, an instance of the input space with $f(\mathbf{z}) = c_z$, $c_z \in \mathcal{Y}$.

Furthermore, let $X = [\mathbf{x}_1, .., \mathbf{x}_n]$ with the corresponding true labels $y_{true} = [y_1, .., y_n]$ be the training set used to train f, and $y_{predicted} = \{f(\mathbf{x}_i) | \mathbf{x}_i \in X\}$ be the predicted labels of the training set. Next, let $\{\mathbf{x} \in \mathcal{X} \ | f(\mathbf{x}) = c_z\}$ and $\{\mathbf{x} \in \mathcal{X} \ | f(\mathbf{x}) \neq c_z\}$ be defined as the *ally* and the *enemy* instances of $\mathbf{z}$ [14], respectively.

LEAFAGE uses X, $y_{predicted}$, $\mathbf{z}$ and c_z to explain why $\mathbf{z}$ was predicted as c_z. It works as follows: *

- A subset of the training set in the neighbourhood of $\mathbf{z}$ is used to build a local linear model. The coefficients of this model provide a measure of importance of each feature locally.
- These coefficients are used to define a local dissimilarity measure between any instance $\mathbf{x_i} \in \mathcal{X}$ and $\mathbf{z}$. In turn, this measure is used to retrieve examples similar to $\mathbf{z}$ from the training set.
- The importance of each feature and the most similar examples are given as explanation of the classification.

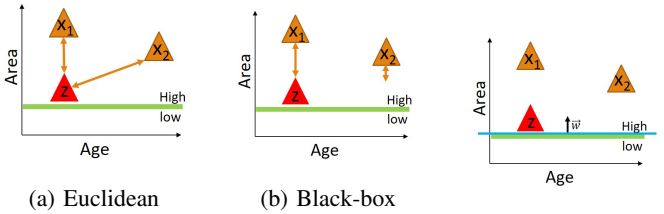
Section III-A and III-B illustrate respectively the adopted dissimilarity measure, and the strategy to build the local linear model. Next, Section III-C explains how LEAFAGE explanations can be presented to the user.

A. *Defining a local dissimilarity measure*

Consider a binary classification problem where an ML model predicts whether a house has a *high* or *low* value according to two features, *area* and *age*, as shown in Figure 2a (in green, the decision boundary of a simple linear classifier). A test house $\mathbf{z}$ is predicted as value *high*. To find similar houses in the training set, one could use the Euclidean distance (Figure 2a). However, this choice does not reflect the reasoning of the classifier, which only looks at the feature *area*; in fact, according to the classifier, $\mathbf{z}$ is more similar to $\mathbf{x}_2$ than $\mathbf{x}_1$ (Figure 2b)

A way to compute a dissimilarity measure that takes into account the reasoning of the classifier is to use feature weights derived from a local linear approximation $\hat{f}_z(\mathbf{x}) = \mathbf{w}_z\mathbf{x} + c$ with $\mathbf{w}_z = (w_{z1}, ..., w_{zd})^T$ of the decision boundary (the blue line in Figure 2c). Then, $\mathbf{w}_z$ will denote the most discriminative direction for the classification of $\mathbf{z}$.

In the simple example of Figure 2 the whole decision boundary can be approximated accurately by a linear model. ML models are usually much more complex, see e.g. Figure 3. However, we assume that locally the closest fragment of the global decision boundary to $\mathbf{z}$ is smooth enough to be linearly approximated (see the blue line in Figure 3).



(a) Euclidean distance: $\mathbf{z}$ is more similar to $\mathbf{x}_1$ than $\mathbf{x}_2$

(b) Black-box classifier: $\mathbf{z}$ is more similar to $\mathbf{x}_2$ than $\mathbf{x}_2$

(c) Approximate the decision boundary with a linear model.

Figure 2: Illustration of different types of distances.

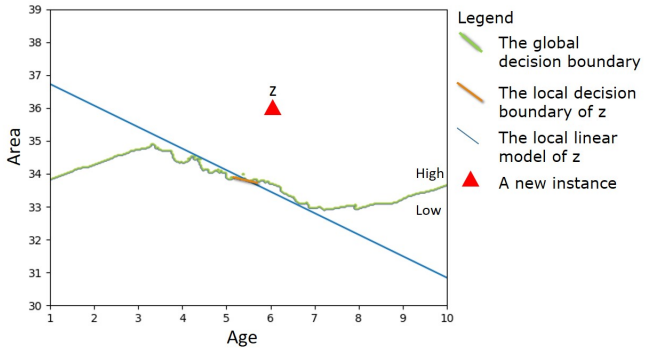


Figure 3: A complex decision boundary that cannot be accurately approximated by a linear model.

The following definitions describe the local behaviour of the ML model around $\mathbf{z}$. These definitions applied to the housing example are illustrated in Figure 3.

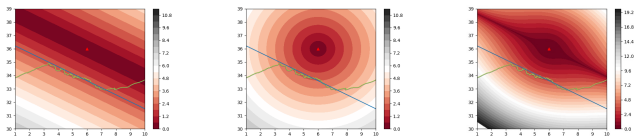
Definition 1. Let *the local decision boundary* of $\mathbf{z}$ be defined as the closest fragment (according to a distance measure $D(\mathbf{x}_1, \mathbf{x}_2)$, e.g. the Euclidean distance) of the global decision boundary to $\mathbf{z}$.

Definition 2. Let *the local linear model* of $\mathbf{z}$ be the model that approximates the local decision boundary of $\mathbf{z}$.

Definition 3. Given the local linear model $\hat{f}_z(\mathbf{x}) = \mathbf{w}_z \cdot \mathbf{x} + c$ of $\mathbf{z}$ let the *the black-box dissimilarity measure* between $\mathbf{z}$ and an instance $\mathbf{t} \in \mathcal{X}$ be defined as the following:

$$b(\mathbf{t}) = D(\mathbf{w}_z^T\mathbf{t}, \mathbf{w}_z^T\mathbf{z}) * D(\mathbf{t}, \mathbf{z}),$$

If D is the Euclidean distance, in the 2D case the black-box dissimilarity has the form shown in Figure 4c. In the first factor of the Definition 3, $\mathbf{w}$ is used as weights to reflect features' importance according to $\hat{f}_z$ (Figure 4a). However, $\hat{f}_z$ is only valid in the neighbourhood N of $\mathbf{z}$, and it is not straightforward to define N. To cope with that, we propose to leverage on the fact that closer instances to $\mathbf{z}$ (according to the D distance measure on the input space) are more likely to be within N: therefore, in Definition 3) a second



(a) First factor

(b) Second factor

(c) Whole formula

Figure 4: Contour-line visualization of the black-box dissimilarity measure.

factor is added, the distance on the input space (Figure 4b). Please note that the dissimilarity measure defined as such, does not always satisfy the condition of identity of indiscernibles $b(\mathbf{z}') = 0 \Leftrightarrow \mathbf{z} = \mathbf{z}'$: depending on $\mathbf{w}_z$, the features that differ between $\mathbf{z}$ and $\mathbf{z}'$ could have no influence on b. Therefore this dissimilarity measure cannot be properly considered a 'metric' in a mathematical sense (instead, it is a 'pseudometric').

B. Computation of the local linear model

The local linear model is computed from a neighborhood of $\mathbf{z}$ sampled from the original training set. Let us denote it as the *local training set* of $\mathbf{z}$.

Methods to sample this local training set have been proposed in LIME [5] and LS [8] (relevant details have been provided in Section II). Both methods have shortcomings related to the right choice of the size of the neighbourhood from which the local training set was sampled.

Taking into account the issues of LIME and LS, we suggest two desired characteristics that a local training set of $\mathbf{z}$ should adhere to:

1) The convex hull of the local training set of $\mathbf{z}$ should contain the local decision boundary of $\mathbf{z}$.
2) There should be enough instances to represent all classes.

We propose a novel sampling strategy that covers both aspects. Its steps are:

1) The local training set of $\mathbf{z}$ is sampled around the local decision boundary of $\mathbf{z}$ (similar to the idea of LS [8]). This makes it possible to sample enough instances from both classes. We assume that the closest enemy $\mathbf{x}_{border}$ of $\mathbf{z}$ from the training set lies close to the local decision boundary of $\mathbf{z}$ and sample around $\mathbf{x}_{border}$.
2) $i_{small} \cdot d$ samples of each class from the training set are sampled, that lie the closest to $\mathbf{x}_{border}$ according to the distance measure D. d instances per class are the minimum amount of examples needed for a good linear approximation, assuming that these d instances lie along the closest decision boundary of $\mathbf{z}$. Since these d instances might not lie exactly along the decision boundary, the amount is increased with i_{small} which is a small integer greater than 1.

An example of this strategy applied on a 2D case with $i_{small} = 10$ is shown on Figure 5. The green and red shapes are instances sampled from the training set to build the local linear model of $\mathbf{z}$.

Given the local training set of $\mathbf{z}$, a linear classification algorithm can be used to build the local linear model of $\mathbf{z}$.

C. Explanation extraction

Given the local linear model $\hat{f}_z(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + ... + w_dx_d$ and an instance $\mathbf{z} = [z_1, .., z_d]$, the importance of each feature z_i can be evaluated as $abs(w_i * z_i)$, and can be provided as an explanation to the user as for which features the original model deems as relevant for its decision on $\mathbf{z}$. We refer to it as *feature importance-based explanation*.

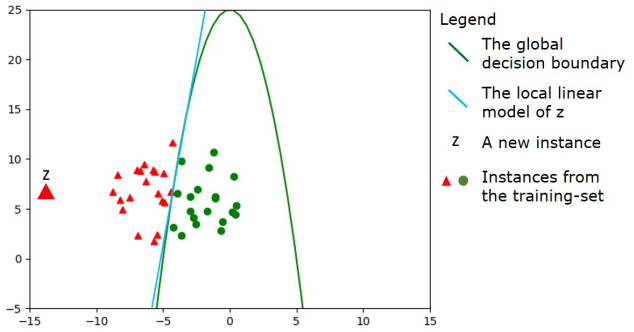


Figure 5: Sampling of the local training set of $\mathbf{z}$.

As discussed in Section II, a way to provide explanations that are closer to how humans think is to use Example-Based Reasoning, i.e. to provide examples that (according to the logic of the black-box model) are related to the test point $\mathbf{z}$. As the logic of the black-box model is locally represented by the black-box dissimilarity measure, the latter can be used to find training examples similar to $\mathbf{z}$ to motivate the decision. Furthermore, one can provide both examples belonging to the predicted class c_z and to the opposite class, which provides insights on the differences between classes according to the black-box model. We refer to them as *example-based explanations*. Feature importance-based and example-based explanations can be also combined to provide better insights.

An example of a test house predicted as *high value* by a black-box model, and a LEAFAGE explanation for this prediction, are shown in Figures 6a and 6b, respectively. The left graph of Figure 6b shows the relative importance of each feature. The two tables on the right show the top 5 similar (according to the black-box dissimilarity measure) houses from the training set, belonging to the same class (*high value*) and from the opposite class *low value*. From these explanations, the user can spot insights on the classification logic, e.g., similar *low value* houses have smaller *living area* than similar *high value* houses.

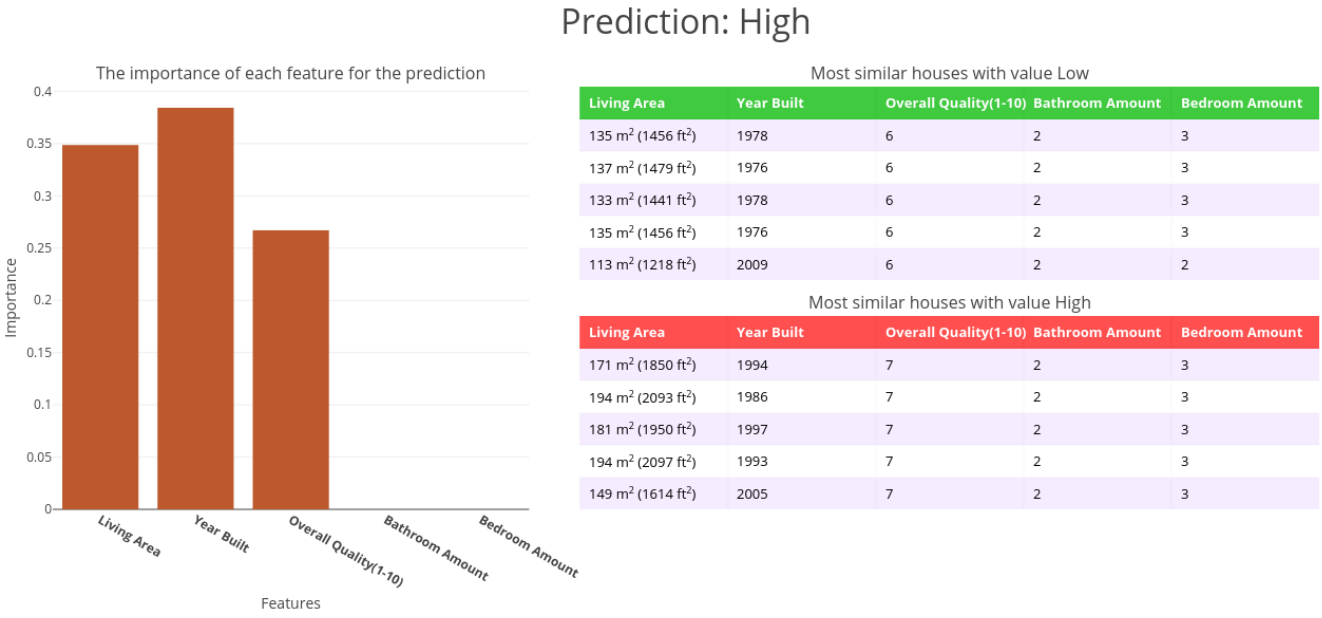
IV. Quantitative Evaluation

This Section evaluates the ability of LEAFAGE to reflect the true local reasoning of a black-box ML model (*faithfulness* of the local approximation).

Four different datasets with different number of features, data points and complexity are used: wine [15], breast cancer [15], banknote [15] and one artificial dataset. The latter is a set of 2D data points from two highly non-separable classes. Instances of each class are sampled from two bi-variate normal distributions with different means ($[0, 0]$ and $[0, 1]$, respectively) and the same covariance matrix ($\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$). The multi-class datasets are converted to binary datasets of one-vs-rest fashion. A combination of a binary dataset and a classifier is referred to as a *setting* in the following. Each dataset is

Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
184 m^2 (1982 ft^2)	1989	7	2	3

(a) A house predicted as value *low* by a black-box model.



Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
135 m^2 (1456 ft^2)	1978	6	2	3
137 m^2 (1479 ft^2)	1976	6	2	3
133 m^2 (1441 ft^2)	1978	6	2	3
135 m^2 (1456 ft^2)	1976	6	2	3
113 m^2 (1218 ft^2)	2009	6	2	2

Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
171 m^2 (1850 ft^2)	1994	7	2	3
194 m^2 (2093 ft^2)	1986	7	2	3
181 m^2 (1950 ft^2)	1997	7	2	3
194 m^2 (2097 ft^2)	1993	7	2	3
149 m^2 (1614 ft^2)	2005	7	2	3

(b) LEAFAGE explanation for the house above.

Figure 6: Example of a LEAFAGE explanation.

randomly split into train (70%) and test set. The train set is used to train six classifiers, namely *Logistic Regression* (LR), *Support Vector Machine* with linear kernel (SVM), *Linear Discriminant Analysis* (LDA), *Random Forest* (RF), *Decision Tree* (DT) and KNN with $K = 1$[1]. In total 36 different settings are tested.

A local linear model $\hat{f}_{x_i}$ is built using LEAFAGE for each instance of the test set. Laugel. et al [8] suggested to test the performance of $\hat{f}_z$ on the *test* instances that fall into a hyper-sphere with a fixed radius and $\mathbf{z}$ as center. Having a fixed radius has a disadvantage that the sphere may include only instances of the same class. Therefore, we propose to use a custom radius by expanding it until the corresponding hyper-sphere includes p percentage of instances that do not have the same predicted label as $c_{\mathbf{z}}$. p should be smaller than, and close to, one (p = 0.95 is used in the experiments), such that the closest testing instances of the opposite class of $\mathbf{z}$ are included and to make the evaluation local, respectively. The scores given by $\hat{f}_z$ are compared with the scores given by the black-box classifier on all the test instances that fall into this hyper-sphere, using the Area Under the ROC (AUC). We define the *average fidelity score* as the average AUC score over the whole test set.

We then compare the average fidelity scores of LEAFAGE (with $i_{small} = 10$) with LIME, in the various experimental settings as shown in Table I.

[1] *scikit-learn 0.19.2* (http://scikit-learn.org/stable/) was used to build these models with their default parameters unless stated otherwise.

Classifier Name	Strategy	Wine			BreastCa.	BankNote	AD
		Class 0 vs rest	Class 1 vs rest	Class 2 vs rest	Benign vs Malignant	0 vs 1	0 vs 1
LDA	LIME	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	99.5 (1.0)	**100 (0.0)**	**100 (0.0)**
	LEAFAGE	100 (0.0)	96.0 (9.4)	100 (0.0)	**99.9 (0.3)**	99.9 (1.7)	98.6 (4.0)
LR	LIME	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	**99.9 (0.6)**	**100 (0.0)**	**100 (0.0)**
	LEAFAGE	**100 (0.0)**	**97.1 (14.2)**	**100 (0.0)**	98.6 (7.8)	99.8 (0.9)	98.6 (4.0)
SVM	LIME	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	**99.9 (0.6)**	**100 (0.0)**	**100 (0.0)**
	LEAFAGE	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	98.6 (7.8)	99.8 (0.9)	99.4 (1.2)
DT	LIME	91.9 (14.9)	**87.9 (22.4)**	91.9 (14.7)	85.0 (16.2)	**99.0 (2.6)**	**59.5 (32.7)**
	LEAFAGE	**92.9 (16.0)**	**85.8 (24.1)**	**100 (0.0)**	**86.5 (18.7)**	**98.7 (4.2)**	**65.0 (33.0)**
RF	LIME	**100 (0.0)**	**99.9 (0.5)**	**100 (0.0)**	**99.9 (0.3)**	**99.1 (2.5)**	61.4 (36.2)
	LEAFAGE	**100 (0.0)**	**99.2 (3.7)**	**100 (0.0)**	**99.9 (0.8)**	**98.7 (3.8)**	**67.4 (32.9)**
KNN	LIME	**98.0 (13.6)**	**62.8 (37.1)**	**60.5 (35.7)**	95.8 (8.2)	**100 (0.0)**	**65.6 (34.3)**
	LEAFAGE	91.3 (15.9)	**62.9 (36.1)**	**60.3 (36.1)**	**97.3 (6.0)**	**99.9 (0.5)**	**65.5 (36.8)**

Table I: Average local fidelity per setting (the standard deviation in brackets). The strategy with the highest mean along with other strategy that are statistically not significantly different are denoted in bold.

Both LIME and LEAFAGE methods perform better than a baseline model (which predicts the majority class), in all settings. Further, both methods work better with linear ML models (SVM, LDA, LR) as opposed as non-linear ones (DT, RF, KNN), especially with the artificial dataset. This was expected, as LEAFAGE and LIME are based on linear

approximations. On linear models, LIME scores significantly better than LEAFAGE in 11 out of 18 settings, while on non-linear models LEAFAGE performs better 5 out of 18 times.

The better performance of LIME on linear ML models could be explained by taking into account that LIME uses a high amount of samples over the whole input space to fit the local linear model. LEAFAGE on the other hand, samples around the closest decision boundary and limits the sampling amount to a minimum (in a sense, it is more *local*). This also explains the better performance of LEAFAGE over LIME on non-linear models.

In conclusion, overall LIME performs better than LEAFAGE on linear ML models, while LEAFAGE performs better on non-linear models.

V. Empirical Evaluation

In order to assess the usefulness of LEAFAGE from the user perspective, we performed a user-study. The target group for this study was the general public. 114 participants with a well spread demographics in term of gender, age and education (but mostly from the Americas) were recruited from Amazon Mechanical Turk, and asked to imagine they were looking for a house to buy, and that they could use an AI application to estimate the value of a house as *low* or *high*. The AI application could also provide an explanation for its estimation.

The IOWA housing dataset [16] was used, from which 5 interpretable features (i.e., features whose meaning can be directly and easily interpreted by humans) of a house were chosen, as shown in Figure 6a. We investigated 4 types of explanations for the prediction of a house, namely *feature importance-based* (Figure 6b left), *example-based* (Figure 6b right), a combination of *example and feature importance-based* (Figure 6b) and *no explanation*, as a baseline.

The evaluation was split into a subjective and an objective part: *perceived aid in decision-making* and *objective transparency*. In the first part, the participants were asked to rate how much they agree to the given explanation from 1 to 5 in terms of: *transparency* (I understand how LEAFAGE made the prediction); *information sufficiency* (the explanation provided has sufficient information to make an informed decision), *competence* (the explanation corresponds to my own decision making) and *confidence* (the explanation made me more confident about my decision). Next, the objective transparency was measured by testing participants as follows: the participants were shown another house, similar to the test one; he/she had to indicate what the system would predict as the sale value of this new house.

Attention checks were implemented; Results were gathered from the 86 participants that passed the checks.

An SVM model with a RBF kernel was trained on the a training set from the IOWA dataset (70%) to predict the binary class (*low* or *high* value). All participants saw forty houses randomly chosen from the test set (30% of the IOWA dataset), with the corresponding predicted value, and one of the four explanation types. All participants saw the same explanations in a randomized order. Finally, the perceived aid in decision making and objective transparency was measured.

	Transparency	Info. Suff.	Competence	Confidence	Objec. Trans.
No Explanation	3.66 (1.03)	3.43 (1.17)	3.70 (0.96)	3.52 (1.1)	8.40 (1.48)
Feature importance	3.92 (0.85)	3.76 (0.97)	3.78 (0.91)	3.68 (1.04)	7.20 (1.66)
Example-based	4.07 (0.76)	4.02 (0.84)	**3.96** (0.86)	**3.98** (0.86)	**8.83** (1.40)
Ex. and Feat.	**4.13** (0.8)	**4.10** (0.83)	3.93 (0.93)	**3.98** (0.9)	8.56 (1.68)

Table II: Results of perceived aid in decision making and objective transparency per explanation type.

Table II shows the results per explanation type and dependent variable. The median score of the explanation types differ significantly over all dependent variables according to Kruskal-Wallis H-tests [17] with $p < 0.001$ and H statistic equal to 124, 202, 55, 125 and 52 (in the left to right order of table II, respectively). The participants perceived getting explanation as more helpful than providing no explanation. Dunn's post-hoc test with Bonferroni correction [18] revealed that in terms of transparency, information sufficiency, competence and confidence both *example-based* and *combination* explanation perform significantly better than *no explanation* and *feature importance-based* explanation, while no significant differences were found between *example-based* and *combination* explanations. Moreover, *feature importance-based* explanation performed significantly better than *no explanation* regarding transparency, information sufficiency and confidence but not in terms of competence. However, in terms of objective transparency, *feature importance-based* explanation performed significantly worse than the rest of the explanation types including *no explanation*. *Example-based* explanation has the highest average objective transparency score, however no statistically significant difference was measured between pairs of *no explanation*, *example-based* and *combination* explanation.

Finally, the participants provided general remarks for each explanation type. When no explanation was provided, they indicated that they could still understand the prediction, but that they needed "complete trust in the system to find it helpful". The participants liked the simplicity and visual aspect of the *feature importance-based* explanation, but they did not find it detailed enough to perform well on the objective transparency part. Moreover, they found it hard to estimate what value of a certain feature changes the prediction, and how the importance really relates to the prediction of a house. Regarding *example-based* explanation, participants appreciated that they could compare similar houses with different sale values. However, some participants disliked this explanation type because of the amount of information present in the tables. Finally, the combination of example-based and feature importance-based explanation received a mixed reaction: some participants liked to get a detailed explanation while others were overwhelmed and focused on one chart and ignored the other.

VI. Conclusion

In this paper, we presented LEAFAGE, a novel method to provide local explanations for the predictions of a black-box ML model. LEAFAGE explanations performed overall better

than the state of the art on non-linear models in terms of local fidelity. We also evaluated LEAFAGE empirically, by engaging people in a user study. This aspect has been largely overlooked by the scientific literature on XAI.

The empirical evaluation showed that overall the participants perceived having explanations behind a prediction as more helpful than having no explanation for the goal of decision making. Interestingly, when participants were tested about their gained knowledge after seeing an explanation, no significant advantage was found compared to providing no explanation. We suspect that this is due to the simplicity of the test, in future work a more comprehensive test could be used to measure the actual transparency. The user study also showed that, with regards to objective transparency, feature importance-based explanations are less effective than providing no explanation at all. This is an important result, which suggests that feature importance-based explanation confuses users more about the prediction than providing no explanation.

Further, example-based explanations performed significantly better than feature-importance based explanation in terms of perceived aid in decision making. Showing both example-based and feature-importance-based explanation did not increase the perceived aid in decision making significantly. This could be due to the overload of information as the participants described. Interestingly, some participants indicated that a tabular view of the example-based explanation was hard to read. In our future work, we will focus on designing them in a more readable and intuitive manner.

ACKNOWLEDGEMENT

This research is supported by the Hybrid AI Explainability and VP AI & Robotics programs at the Netherlands Organisation for Applied Scientific Research (TNO).

REFERENCES

[1] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1):39–59, 1994.

[2] Shahina Begum, Mobyen Uddin Ahmed, Peter Funk, Ning Xiong, and Mia Folke. Case-based reasoning systems in the health sciences: a survey of recent trends and developments. *IEEE Trans. on Sys., Man, and Cyb., Part C (Applications and Reviews)*, 41(4):421–434, 2011.

[3] Isabelle Bichindaritz and Cindy Marling. Case-based reas. in the health sci.: What's next? *Arti. intel. in medic.*, 36(2):127–135, 2006.

[4] Janet L Kolodner. An introduction to case-based reasoning. *Artificial intelligence review*, 6(1):3–34, 1992.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Expl. the predic. of any class. In *ACM SIGKDD Int. Conf. on Know. Disc. and Data Min.*, pages 1135–1144. ACM, 2016.

[6] P.H.N. Gill. *Intr. to ML Interpre.* O'Reilly Media, Incor., 2018.

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Int. Conference on Mach. Learning*, pages 2048–2057, 2015.

[8] Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretablity. *arXiv preprint arXiv:1806.07498*, 2018.

[9] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 36, 2017.

[10] Michael M Richter and Rosina O Weber. *CBR*. Springer, 2016.

[11] Nava Tintarev and Judith Masthoff. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*, pages 479–510. Springer, 2011.

[12] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4):367–382, 2014.

[13] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455, 2008.

[14] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Comparison-based inverse classification for interpretability in machine learning. In *Int. Conf. on Inf. Proc. and Manag. of Uncert. in KB Systems*, pages 100–111. Springer, 2018.

[15] Dua Dheeru and Efi Karra Taniskidou. UCI ml repository, 2017.

[16] Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regr. project. *Jour. of Stat. Edu.*, 19(3), 2011.

[17] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[18] Olive Jean Dunn. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252, 1964.