Example and Feature importance-based Explanations for Black-box Machine Learning Models

Ajaya Adhikari[1], D.M.J. Tax[2], Riccarod Satta[1], and Matthias Fath[1]

[1]TNO, The Hague
[2]Delft University of Technology, Delft

December 24, 2018

Abstract

As machine learning models become more accurate, they typically become more complex and uninterpretable by humans. The black-box character of these models holds back its acceptance in practice, especially in high-risk domains where the consequences of failure could be catastrophic such as healthcare or defense. Providing understandable and useful explanations behind ML models or predictions can increase the trust of the user. Example-based reasoning, which entails leveraging previous experience with analogous tasks to make a decision, is a well known strategy for problem solving and justification. This work presents a new explanation extraction method called LEAFAGE, for a prediction made by any black-box ML model. The explanation consists of the visualization of similar examples from the training set and the importance of each feature. Moreover, these explanations are contrastive which aims to take the expectations of the user into account. LEAFAGE is evaluated in terms of fidelity to the underlying black-box model and usefulness to the user. The results showed that LEAFAGE performs overall better than the current state-of-the-art method LIME in terms of fidelity, on ML models with non-linear decision boundary. A user-study was conducted which focused on revealing the differences between example-based and feature importance-based explanations. It showed that example-based explanations performed significantly better than feature importance-based explanation, in terms of perceived transparency, information sufficiency, competence and confidence. Counter-intuitively, when the gained knowledge of the participants was tested, it showed that they learned less about the black-box model after seeing a feature importance-based explanation than seeing no explanation at all. The participants found feature importance-based explanation vague and hard to generalize it to other instances.

1 Introduction

Machine Learning (ML) is a rapidly growing field. There has been a surge of complex black-box models with high performance. On the other hand, the application of these models especially in high-risk domains is more stagnant due to lack of transparency and trust in these black-box models. There is a disconnect between the black-box character of these models and the needs of the users. A sub-field of eXplainable Artificial Intelligence (XAI) has emerged to fix this disconnect but it is still in its early stages.

Case-based or example-based reasoning (i.e. motivating a decision by providing examples of similar situations, that led to same decision in the past) is widely recognized as an effective way to provide explanations [1], as it bears a close resemblance to the way humans naturally think. As a result, it is commonly used in the health-care sector for decision-support systems [2, 3] and in law for justifying arguments, positions and decisions [11]. However, the usage of Case-Based Reasoning (CBR) to explain decisions taken by a particular category of such support systems, i.e. those using black-box Machine Learning (ML) models (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. Yet, black-box ML models are

becoming widespread, as they usually outperform transparent models by a large margin [22]. Notably, most of the scientific literature on explainable ML for black-box models focuses instead on evaluating the importance of the single features used for the decision (feature importance-based explanations, see e.g. LIME [23]).

In this paper, we propose a new method for providing CBR explanations of the local reasoning of black-box models. Here, *local* refers to the ability of tailoring the explanation to a single decision/prediction taken by the ML model, as opposed to providing a global explanation of how the whole model works in general. 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. LEAFAGE supports both direct explanations (similar examples supporting the prediction made by the ML model) and contrastive explanations (similar counter-examples that have a different prediction). Contrastive explanations are generally considered as quite effective by social scientists [15, 17]. Through a contrastive explanation the user will be able to understand why a data point was predicted as class A instead of another class B.

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. It is evaluated by comparing the predictions made by the linear model (which was used to extract the explanation) with the predictions made by the underlying black-box model, in the neighbourhood of the instance being explained. The *usefulness* to the user is evaluated by conducting a user-study in terms of perceived aid in decision-making, measured transparency and persuasion. More specifically, we focus on the comparison between example-based and feature importance-based explanation.

The remainder of this paper is structured as follows. In Chapter 2, we present background information and related work on XAI, and explore best approaches to provide explanation by leveraging on social research. Chapter 3 describes LEAFAGE, which is then evaluated in terms of fidelity and usefulness to the user, respectively in Chapter 4 and 5. Finally, Chapter 6 draws conclusions and possible future research directions.

2 Background

2.1 Explanations in machine learning

This Section explores the problem of extracting explanations from ML models. A taxonomy of the available methods is first presented in Sections 2.1.1 and 2.1.2. Then, in Section 2.1.3 we focus on existing work that is more directly related to our method.

2.1.1 Local vs Global explanation

An explanation about a ML model can be of *global* or *local* scale. A *global* explanation clarifies the inner workings of a whole ML model, or how the relationship between input and output spaces is modeled [9]. *Local* explanations instead look at the reasoning behind a single prediction, thus targeting a sub-region of the input space. As the complexity of the ML model grows, it become harder to generate an understandable global explanation. In such cases, a model is practically a black-box [9]. Instead, 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.

An example of a decision tree model is shown in figure 1a. The classification problem is to classify a house as value *low* or *high*, given its *age* and *area*. The whole decision tree can be seen as a global explanation, because it explains how the features *age* and *area* of a house is used to determine its value. But if one is only interested to understand why a particular house is predicted with a certain value a local explanation is needed. For example a house with *age*=10 and *area*=30 is predicted as value *high* because its *age* is greater than 5 and its *area* is greater than 20.

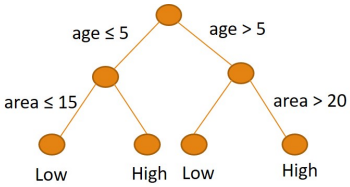
2.1.2 Explanation extraction strategy

In the literature there are three main strategies that extract human-understandable explanations from ML models, namely globally transparent, model-oriented and model-agnostic. In the first strategy, the ML model is optimized to be accurate but also simple enough to be understandable by humans. While in the second and third strategy, the ML model is not required to be globally transparent, instead human-understandable explanations are extracted from the complex ML model. The difference between the last two strategies lies in the type of information used for the explanation extraction.

In high-risk applications such as health-care and defense, globally transparent and interpretable model are preferred because it is important to understand how the ML model exactly works to avoid catastrophic consequences. The most prominent ML models that are regarded as transparent and interpretable in the literature, are decision trees, decision rules and linear models [10]. These models have two drawbacks. Firstly, they are not inherently transparent and interpretable [16], despite their apparent simplicity. The size of these models can be enormous, making them complex and hard to grasp e.g. a decision tree model containing a lot of nodes or a long decision rules model or a linear model in a high dimensional space. They could be restricted to have a small size, but that can reduce the performance of these models significantly. Secondly, in certain domains complexer ML models such as deep neural-networks have proven to perform better than transparent models. In that case, choosing a transparent model will be at the cost of its performance.

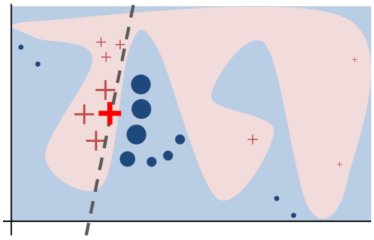
In model-oriented strategy, the internal workings of the ML model is leveraged to extract an explanation. For example, Xu et al. created an explanation method that can explain why the machine learning model classified certain contents in an image. The explanation consists of an attention map that shows which part of the image was important for the prediction. Figure 1b shows two examples, in which a mask placed over the original images shows which parts of the images were important for the prediction of a frisbee and a dog. The inner-workings of a LSTM ML model is directly leveraged to build the attention map. This strategy has the advantage that the explanations reflect the true reasoning of the ML model with high accuracy. The downside is that the explanation methods are model-dependent, requiring a different explanation method for each type of ML model.

In model-agnostic strategy, the ML model is seen as a black-box. No internal information about the ML model is used, rather the black-box model is queried using a set of instances from the input space. The outputs of these queries are used to gain insights in the behaviour of the model, given the inputs. In some cases the training instances are used to query the black-box model [14] and in other cases new generated instances [23]. An advantage of a model-agnostic explanation method is that it can be used for any type of ML model. Further, the ML model can be built without taking explanability into account. A downside of these methods is that they can be less accurate than the other strategies in terms of reflecting the true reasoning of the ML model. LIME [23] is a well known model-agnostic method. LEAFAGE uses the basic ideas of LIME, hence it is important to have a good overview of LIME. In section 2.1.3 LIME along with another related method LS [14] is explained.



(a) Transparent model strategy e.g. a decision tree.

(b) Model oriented strategy: this explanation leverages information from the underlying model [26].



(c) Model agnostic strategy e.g. LIME [23]

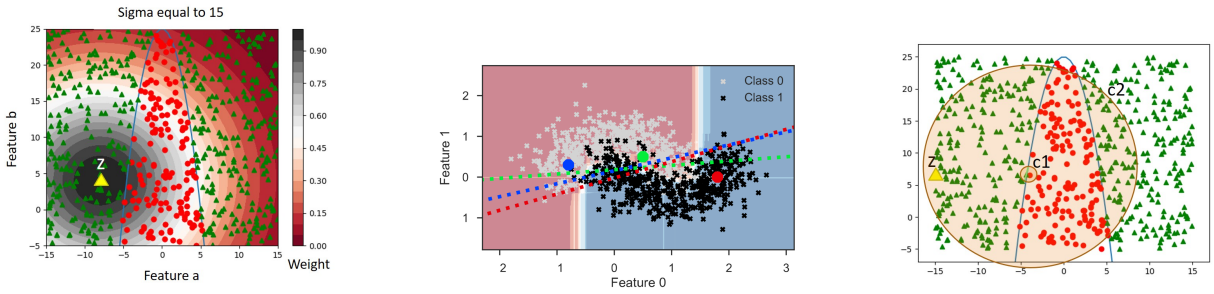
Figure 1: Examples of different types of strategies for explanation extraction.

2.1.3 Local linear approximations

LIME [23] stands for Local Interpretable Model-agnostic Explanations. Given the prediction of an instance $\mathbf{z}$ by a black-box classifier, LIME provides a local explanation which states why this prediction was made. It focuses on a small neighbourhood which is centered around $\mathbf{z}$. LIME approximates the decision boundary of the black-box classifier in this neighbourhood by a linear model and extracts an explanation from it.

Figure 1c shows the main idea behind LIME. The blue/red background represents the black-box decision function. The bold plus point $\mathbf{z}$ is predicted as red and needs an explanation for its prediction. Artificial instances Z (other plus and minus points) are sampled from a distribution modelled as uni-variate normal distributions of each feature in the training-set. These instances are given weights w (represented by the size) according to their proximity to $\mathbf{z}$, by using an exponential kernel. The hyper-parameter σ of the kernel determines how fast the proximity value decreases when going further from $\mathbf{z}$, i.e. it determines the size of the neighbourhood of $\mathbf{z}$. Next, the black-box predictions y of these Z instances are retrieved and used to train a linear model g (the dashed line). The weights w are used in the optimization of g, such that high importance is given in classifying Z instances close to $\mathbf{z}$ correctly, than other Z instances far away. At last, the importance of each input feature is extracted from g.

In LIME the size of the neighbourhood, that is used to estimate the local linear model, is determined by σ and in its implementation it is set to $0.75 * \sqrt{d}$ (d is dimension of the input space). An example is given in figure 2a, in which the neighbourhood is too big which leads to a linear approximation of the global parabola decision boundary (over-generalization). It is also possible that the neighbourhood is too small and does not include the decision boundary close to $\mathbf{z}$. This suggest that it is important to tune the size of the neighbourhood according to how far $\mathbf{z}$ is to the closest decision boundary.



(a) A bad sampling scenario of LIME.

(b) Limitation of LIME spotted by Laugel et al. [14]

(c) Two bad sampling scenario of LS [14]

Figure 2: Limitations in the sampling strategies of LIME and LS

Laugel et al. [14] spotted the problem of over-generalization of the local decision boundary by LIME. They show an example (figure 2b) in which the linear approximations (dotted lines) learned by LIME for the prediction of the 3 biggest colored points are shown. The black-box boundary is given by the blue and pink background. The linear models of the points are very similar even-though the local decision boundaries are different.

Laugel et al. [14] propose a new method LS (Local Surrogate) as an improvement on LIME and suggest to sample instances from the training-set within a hyper-sphere with the center equal to the closest decision boundary of $\mathbf{z}$, and a radius r. r is fixed according to the maximum distance between the training instances. The closest decision boundary of $\mathbf{z}$ is approximated by using the growing spheres algorithm [13], in which artificial instances are sampled uniformly within a hyper-sphere of growing radius centered on $\mathbf{z}$, until an instance from the opposite class $\mathbf{z_{border}}$ is found. The $\mathbf{z_{border}}$ is regarded as the approximation of the local decision boundary. This brute-force way of approximating the local decision boundary accurately is only feasible in a small dimensional input space but not in a high one, because the volume of hyper-sphere

increases exponentially with the number of dimensions. Like LIME, LS also has two sampling scenarios (figure 2c), in which the neighbourhood can be too small (circle $c1$) or too big (circle $c2$).

2.2 Users' perspective on explanation

XAI research has been criticized for creating explanations only from a technical point of view and mostly ignoring the real usefulness of the explanations to the end-users [18]. Different social science research fields have investigated how people communicate in real life and what people expect from an explanation [17]. In this study, we leverage two findings from social science research namely CBR and contrastive explanations. Further, we look into different evaluation methods of explanations from the perspective of the user.

In CBR the reasoner relates to previous experiences to understand and solve a current problem he/she faces [11, 24]. This type of reasoning lies very close to how we as humans think [3, 1]. We use it in our daily lives to solve problems. The process of CBR can be typically divided into three steps namely retrieve, adapt and learn. For example, let's take a problem of choosing a dish to cook in this situation: no onions at home and wanting a light meal. We might think back to the different previous dishes we made and choose a few of them that suit the current situation (retrieve step). It is possible that the chosen dishes do not completely fulfill the requirements of the current situation e.g. all the known dishes use onions. In that case the most suitable known dish is picked and adapted (adapt step) e.g. use cheese instead of onions. The new dish can be leveraged to make future dishes or avoided completely (learn step), according respectively to the liking or disliking of the dish.

The applications of CBR can be divided into two types namely problem-solving and decision-justification [11]. In problem-solving previous similar situations are used as aid to decide how to proceed with the current situation. While in decision-justification previous similar situations are leveraged to support or dismiss certain possible decisions. CBR for the purpose of problem-solving is commonly used in the health-care sector [3, 2]. For example, physicians think back to patients from the past that had similar symptoms as the patient they are examining. They remember the diagnosis of those previous patients and which treatment worked. That information helps them to diagnose the patient in front of them and to suggest a treatment. Moreover, real examples of medical cases in past are being used to train health-care professionals, complete guidelines and provide anecdotal accounts of treatments of individual patients in the medical literature [3]. Furthermore, CBR software systems are being used to aid health care professionals in retrieving similar medical cases from the past [2]. Law is a prominent domain in which CBR is used for the goal of justifying arguments, positions and decisions. Layers use CBR to justify a position by providing supporting relevant cases from the past [11]. Moreover, common law, which is widely used in most English-speaking countries, is based on precedence i.e. judicial decision made on similar cases from the past [24].

Another finding states that generally people ask contrastive questions. When people ask a why-question often they do not want the whole explanation rather they are interested in a subset that can answer the conflict between the observed event and their own mental model of causation. Miller et al. [17] states "people do not ask why event P happened, but rather why event P happened instead of some event Q", which are called contrastive questions. Most researchers in psychology, philosophy and cognitive science agree that all why-questions are contrastive [17]. We use Lipton's [15] definition of the events P and Q being respectively the *fact* and *foil*. The fact is the event that occurred and the foil is the event that did not occur. In machine learning context, a contrastive explanation answers why class A (the fact class) was predicted by the machine learning model instead of class B (the foil class).

At last, the goal of an explanation system of a ML model is to be useful in practice for the intended users, hence conducting user-studies on the explanation system is important. 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 [25, 21, 8, 4, 20]. Before conducting a user-study it is important to understand what goals the explanation system tries to achieve. Tintarev [25] defines seven goals for an explanation system, namely transparency, scrutability, trust, effectiveness, efficiency, persuasiveness and satisfaction. All of the goals can be evaluated subjectively by asking for the

opinion of the user [25]. Moreover, trust and satisfaction are subjective by nature, and can only by evaluated subjectively. But, transparency, effectiveness, efficiency and persuasiveness can also be evaluated/measured objectively, by testing the users whether they have understood the reasoning behind the recommendations [23], evaluating whether the users get better suited recommendations with explanations compared to without any explanations [4], measuring the interaction time [8] and checking whether the user buys an recommended item, respectively.

3 LEAFAGE

We propose a new method LEAFAGE that provides intuitive and understandable explanations for a prediction made by any black-box ML model. LEAFAGE stands for Local Example and Feature importance-based model AGnostic Explanation. The explanation makes the reasons behind a prediction transparent to the user, by providing examples from the training-set that are similar to the instance being explained and showing 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 $\mathcal{Y} = \{c_1, c_2\}$. Let $\mathbf{z} \in \mathcal{X}$ be an instance of the input space with $f(\mathbf{z}) = c_z$, $c_z \in \mathcal{Y}$, for which an explanation is needed. Furthermore, let $X = [\mathbf{x}_1, .., \mathbf{x}_n]$ with the corresponding true labels $y_{true} = [y_1, .., y_n]$ be the training-set that f was trained on. Let $y_{predicted} = \{f(\mathbf{x}) | \mathbf{x_i} \in X\}$ be the predicted labels of the training-set. Next, let $\{\mathbf{x} \in \mathcal{X} \mid f(\mathbf{x}) = c_z\}$ and $\{\mathbf{x} \in \mathcal{X} \mid f(\mathbf{x}) \neq c_z\}$ be defined as the ally and the enemy instances of $\mathbf{z}$ [13], respectively. LEAFAGE needs X, $y_{predicted}$, $\mathbf{z}$ and c_z to extract an explanation for why $\mathbf{z}$ was predicted as c_z instead of the opposite class.

In a high-level overview, LEAFAGE works as follows. First, a sub-set of the training-set in the neighbourhood of $\mathbf{z}$, is used to build a local linear model, that approximates f in the neighbourhood of $\mathbf{z}$. The importance of each feature for the classification of $\mathbf{z}$ is extracted from that linear model. Next, the importance of each feature is used to define a similarity measure that returns how similar an instance $\mathbf{x} \in \mathcal{X}$ is to $\mathbf{z}$. This similarity measure is used to retrieve similar examples as $\mathbf{z}$ from the training-set. At last, the importance of each feature along with the most similar examples are given as an explanation for why f classified the instance $\mathbf{z}$ as c_z.

In subsections 3.1 and 3.2 the similarity measure and the strategy to build the local linear model are defined, respectively. Further, subsections 3.3 and 3.4 explain how an LEAFAGE explanation can be made contrastive and visualized, respectively.

3.1 Defining Similarity

A classification example is shown in figure 3, in which the black-box classifier predicts whether a house has a *high* or *low* value according to its *area* and *age*. A new house $\mathbf{z}$ is predicted as value *high*. To find similar houses a similarity measure has to be defined. A trivial solution is to use euclidean distance measure which gives equal weights to all features. Figure 3a shows two potential houses ($\mathbf{x}_1$ and $\mathbf{x}_2$) from the training-set. According to the euclidean distance, $\mathbf{z}$ is more similar to house $\mathbf{x}_1$ than house $\mathbf{x}_2$. But the black-box classifier only looks at the feature *area*. Thus, according to the black-box classifier $\mathbf{z}$ is more similar to $\mathbf{x}_2$ than $\mathbf{x}_1$ (figure 3b)

The importance of each features for the classification of $\mathbf{z}$ can be retrieved by approximating the decision boundary of the black-box classifier linearly as shown by the blue line in figure 3c. Let $g(\mathbf{x}) = \mathbf{w}_z\mathbf{x} + c$ with $\mathbf{w}_z = (w_{z1}, ..., w_{zd})^T$ be the linear model that approximates the decision boundary. $\mathbf{w}_z$ denotes the most discriminative direction for the classification of $\mathbf{z}$.

In figure 3 the global decision boundary is a horizontal line which can be approximated accurately by a linear model. But in practice the decision boundary of the black-box ML model can be arbitrarily complex as shown in figure 4. In that case, a linear model will not be able to approximate the the global decision boundary accurately. However, we assume that a small fragment of the global decision boundary which is

the closest to $\mathbf{z}$ is smooth enough to be linearly approximated accurately, as illustrated by the blue line in figure 4. In that case the linear model is not valid globally, but only locally in the neighbourhood of $\mathbf{z}$. We further assume that the closest decision boundary to $\mathbf{z}$ is the most important fragment of the global decision boundary for the classification of $\mathbf{z}$ [14, 13].

The following definitions describe the local behaviour of the black-box ML model around $\mathbf{z}$. These definitions applied to the housing example are illustrated in figure 4.

Definition 1. Let *the local decision boundary* of $\mathbf{z}$ be defined as the closest fragment (according to euclidean distance) of the global decision boundary to $\mathbf{z}$.

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

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

$$b(\mathbf{t}) = \sqrt{d} \cdot \left\| \mathbf{w}_z^T \mathbf{t} - \mathbf{w}_z^T \mathbf{z} \right\| + \|\mathbf{t} - \mathbf{z}\|,$$

The black-box similarity values on the two dimensional example is shown on figure 5c. In the first term of the black-box similarity formula, $\mathbf{z}$ and $\mathbf{t}$ are projected onto the direction $\mathbf{w}$ (extracted from g) and euclidean distance is applied onto those projected values, because $\mathbf{w}$ it is the most discriminative direction for the classification of $\mathbf{z}$. But g is only valid in the neighbourhood N of $\mathbf{z}$, and it is not straightforward to define N. We propose a heuristic solution in which the fact that closer instances to $\mathbf{z}$ (according to the euclidean distance on the input space) are more likely to be within N, is leveraged. The black-box similarity measure (definition 3) weights the euclidean distance on the input space (5a) and the euclidean distance on the vector $\mathbf{w}$ (5b) equally.

3.2 Computation of the local-linear model

Definition 4. Let *the local training-set* of $\mathbf{z}$ be defined as the instances $X_z = \mathbf{x}_1, .., \mathbf{x}_t$, $\forall \mathbf{x} \in \mathcal{X}$ with labels $y_l = \{f(\mathbf{x}) | \mathbf{x} \in X_z\}$, which are used to build the local linear model of $\mathbf{z}$.

For the generation an accurate local-linear model the sampling strategy of the local training-set of $\mathbf{z}$ is very important. Methods LIME [23] and LS [14] have proposed solutions for sampling instances, which are explained and reviewed in detail in section 2.1.3. Their shortcomings were related to the size of the chosen neighbourhood from which the local training-set was sampled. We suggest two desired characteristics that a local training-set of $\mathbf{z}$ should adhere to, taking the shortcomings of LIME and LS into account:

1. The convex hull of the local training-set of $\mathbf{z}$ should contain the local decision boundary of $\mathbf{z}$.

(a) According to the euclidean distance $\mathbf{z}$ is more similar to $z1$ than $z2$

(b) According to the black-box classifier $\mathbf{z}$ is more similar to $z2$ than $z1$

(c) The black-box global decision boundary can be approximated accurately by a linear model.

Figure 3: Simple decision boundary of a black-box ML model that prediction whether a house has *low* or *high* value. A new house $\mathbf{z}$ is predicted as value *high*.

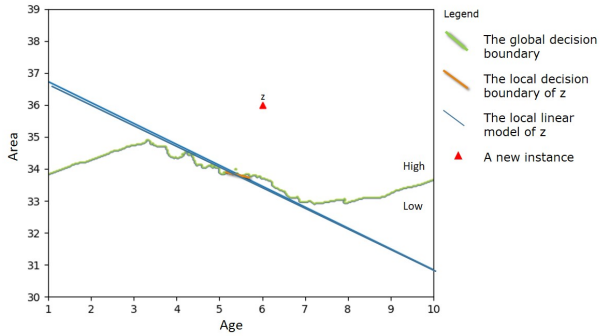
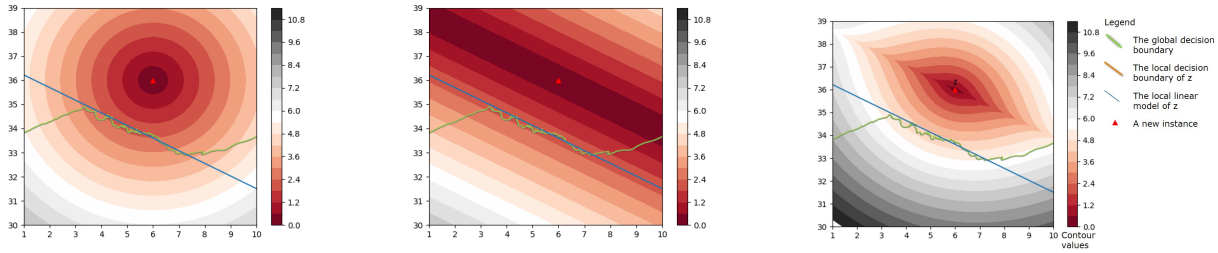


Figure 4: A complex decision boundary that cannot be accurately approximated by a linear model.



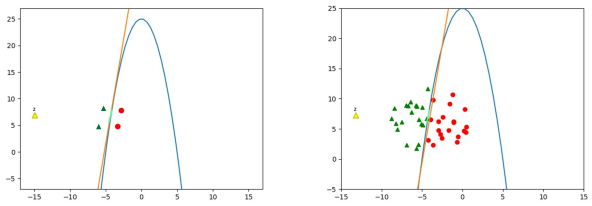
(a) Contour-lines of the first term of the similarity measure.

(b) Contour-lines of the second term of the similarity measure.

(c) Contour-lines of the black-box similarity measure.

Figure 5: An example to illustrate the black-box similarity measure.

2. There should be enough instances of both classes. The minimum amount of instances per class that lie along the local decision boundary, which are needed for a good linear approximation is equal to the dimension d of the input space. For example in a two dimensional space, two instances from each class that lie along the local decision boundary of $\mathbf{z}$ are needed as shown in figure 6a.



(a) The minimum amount of instances needed.

(b) Sampling of the local training-set of $\mathbf{z}$.

Figure 6: An example to illustrate the linear approximation of the local decision boundary of $\mathbf{z}$.

Given the desired characteristics of a local training-set we propose a new sampling strategy. Its two steps and motivations are listed below:

1. The local training set $\mathbf{z}$ is sampled around the local decision boundary of $\mathbf{z}$ (similar to the idea of LS [14]). This makes it possible to sample enough instances from both classes avoiding a bad sampling scenario of LIME (figure 2a). 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 $\mathbf{x}_{border}$. The growing spheres algorithm (see section 2.1.3) suggested by LS is not used, because it is not accurate in a high dimensional input space.

2. LS had two bad sampling scenarios (figure 2c) in which there were too few or too many local training instances, that lead to bad linear approximations. To avoid these scenarios as much as possible, this method proposes to sample $i_{small} \cdot d$ examples of each class from the training-set that lie the closest to $\mathbf{x}_{border}$. The number of samples per class is dependent on the input dimension d, because 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 n. These d instances might not lie exactly along the decision boundary, thus the amount is increased with i_{small} which is a small integer greater than 1. This strategy applied on a two dimensional example with $i_{small} = 10$ is showed on figure 6b. The green and red shapes are instances sampled from the training-set to build the local linear model of $\mathbf{z}$. The local linear model of $\mathbf{z}$ is able to approximate the local decision boundary of $\mathbf{z}$ accurately.

At last, given the local training-set of $\mathbf{z}$, any linear classification algorithm can be used to build the local

linear model of $\mathbf{z}$.

Algorithm 1: *BIS(*$\mathbf{z}$, c_z, X, y, i*)*

Input : $\mathbf{z}$: instance to explain prediction of, c_z: predicted label of $\mathbf{z}$, X:the training-set, y:the predicted labels of $\mathbf{z}$, i_{small}:the amount of instances per dimension and class to sample (default value:3)

Output: The local training-set of $\mathbf{z}$ (X_z) and the corresponding labels y_n

$X_a \leftarrow \{\mathbf{x}_i \in X \mid y_i = c_z\}$ `// the allies of n from the training-set`
$X_e \leftarrow \{\mathbf{x}_i \in X \mid y_i \neq c_z\}$ `// the enemies of n from the training-set`
$\mathbf{x}_{border} \leftarrow \arg\min \|z - x\|,\ x \in X_e$
$amount \leftarrow i_{small} \cdot d$ `// the amount of instances to sample per class`
$d_a = \{\|\mathbf{x}_{border} - \mathbf{x}\| \mid x \in X_a\}$
$d_e = \{\|\mathbf{x}_{border} - \mathbf{x}\| \mid \mathbf{x} \in X_e\}$
$X'_a \leftarrow \text{sort}(X_a,\ key = d_a)[1\ ..\ amount]$
$X'_e \leftarrow \text{sort}(X_e,\ key = d_e)[1\ ..\ amount]$
$X_z \leftarrow \text{concatenate}(X'_a,\ X'_e)$
$y_z \leftarrow$ corresponding y values of X_z
return X_z, y_z

3.3 Generating contrastive explanations

One of the important finding from social research is that *why questions* are contrastive. In ML context, if a ML model predicts an instance $\mathbf{z}$ to be of class $c_\mathbf{z}$ than the user might ask "Why did the ML model predict this instances as class $c_\mathbf{z}$ instead of class c_f". The user expected class c_f and wants a specific explanation of why $c_\mathbf{z}$ (the fact class) was predicted instead of c_f (the foil class). If the ML model solves a binary classification, then any explanation is by definition contrastive and the methods describes in the previous sections can be used. On the other hand if the ML model solves a multi-class problem, providing a contrastive explanation is not trivial.

First, the foil class has to be determined. Determining the foil class means understanding which class the user expected instead of the fact class. Performing this task exactly is a very hard task. We propose a heuristic method to get the foil class. Generally, the ML model gives a score for each possible class. A score of a class states how likely the instance $\mathbf{z}$ belongs to that class. The predicted class $c_\mathbf{z}$ has the highest score. We regard the class with the second highest score as the foil class c_f.

The instances from the training-set that have the fact or the foil class as predictions are selected. With this filtered training-set along with the predicted labels, a LEAFAGE explanation can be extracted that explains why the fact class was predicted instead of the foil class.

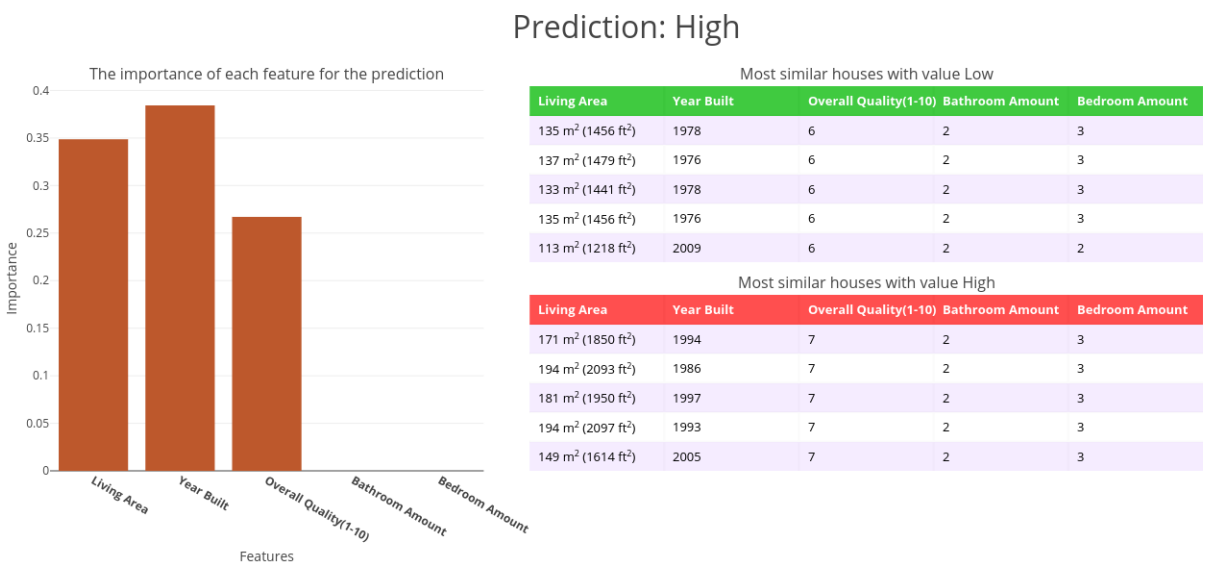
3.4 Explanation extraction

Given the local linear model $g(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + ... + w_dx_d$ of an instance $\mathbf{z} = [z_1, .., z_d]$, the importance of each feature for the prediction of $\mathbf{z}$ can be extracted. The prediction of $\mathbf{z}$, can be determined by the score $g(\mathbf{z})$. Thus, the contribution of a feature value z_i, to the prediction is equal to $abs(w_i * z_i)$. The magnitude of this contribution value denotes the importance of the feature value z_i to the prediction. The importance of each feature value for the prediction can be provided as an explanation to the user. Moreover, a small subset of feature values with high importance can be chosen to give as an explanation.

Social research indicated that Example Based Reasoning lies very close to how we as humans think. It can be used for problem-solving and decision-justification. Given the black-box similarity measure b of $\mathbf{z}$, instances that have similar classification logic can be extracted from the training-set. Furthermore, one can provide similar instances from the fact class and the foil class. That can give insights on the differences between the instances from the fact and foil class in the neighborhood of $\mathbf{z}$.

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

Figure 7: Example of a house that is predicted as value low by the machine learning 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

Figure 8: Example of a LEAFAGE explanation.

We propose to combine feature importance-based explanation with example-based explanation. This combination will allow the user to understand the importance of each feature value and the differences between the instances from the fact and foil class. An example of a house that is predicted as value low by a ML model and an LEAFAGE explanation for its prediction are shown in figures 7 and 8, respectively. The left graph of figure 8, shows which of the feature values of the house were the most important to make the prediction. The length of each bar shows the relative importance of each feature. In this case, the feature values *Bathroom Amount=2* and *Bedroom Amount=3* are not important. Feature value Year *Built=1989* is the most important followed by *Living Area=184*m^2 and *Overall Quality=7*. The two tables in the right show houses that are similar to the house being explained. The green table shows 5 similar houses that have low value and the red table shows 5 similar houses with high value. These tables make clear how big the differences for each feature are, between similar houses with value low and high. For example in this explanation, the differences in Living Area are big between low and high value houses while there are no differences in Bathroom Amount and Bedroom Amount.

4 Quantitative Evaluation

This section evaluates the ability of LEAFAGE in reflecting the true local reasoning of the underlying black-box ML model. It evaluates how faithful the local linear model is to the underlying black-box model, because the local linear model is used to extract both example and feature-importance-based explanations. To demonstrate the model-agnostic character of LEAFAGE, the evaluation is done on a combination of six different types of black-box ML models and five different datasets (four UCI [6] datasets and one artificial dataset). It compares the fidelity of LEAFAGE, LIME [23] and a baseline model, with each other.

4.1 Evaluation method

A dataset is first split into two disjoint sets of training-set (X_{train}, y_{train}) and testing-set (X_{test}, y_{test}), which contain 70% and 30% of data, respectively. The training-set is used to build a black-box classifier f. Further, a local linear model is built for each instance of the testing-set. To build these local linear models, only the training-set and the black-box classifier f are available. The testing instances X_{test} with their predicted labels $y_{test\_f} = \{f(\mathbf{x}_i)|\mathbf{x}_i \in X_{test}\}$ are used to get a *local fidelity score* of each local linear model. At last, the individual local fidelity score of all the local linear models are averaged to get one fidelity score per dataset and classifier. We propose a new method of getting the local fidelity score of a local linear model $g_{\mathbf{z}}$ of an instance $\mathbf{z}$ with $f(\mathbf{z}) = c_{\mathbf{z}}$

The local linear model should be valid in the neighbourhood of $\mathbf{z}$. The difficulty lies in how to set the size of this neighbourhood. Method LS [14] suggested to test the performance of $g_{\mathbf{z}}$ on the testing instances that fall into a hyper-sphere with a fixed radius and $\mathbf{z}$ as center. Having a fixed radius has a disadvantage that it may include only instances of the same class. We propose a custom radius per $g_{\mathbf{z}}$. The hyper-sphere with center $\mathbf{z}$ is expanded until it 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 z are included and to make the evaluation local, respectively. The test instances $X_{test\_\mathbf{z}}$ that fall into this hyper-sphere are used to get the fidelity score of $g_{\mathbf{z}}$. The labels given by the black-box classifier f are compared to the scores given by the local linear model $g_{\mathbf{z}}$, using the AUROC evaluation metric. The local fidelity score of a local linear model $g_{\mathbf{z}}$ of instance $\mathbf{z}$, to a black-box classifier f is defined as follows:

$$AUC(\{f(\mathbf{x}_i)|\mathbf{x}_i \in X_{test\_\mathbf{z}}\}, \{g_{\mathbf{z}}(\mathbf{x}_i)|\mathbf{x}_i \in X_{test\_\mathbf{z}}\})$$

The average local fidelity scores of three strategies are compared per dataset and classifier. These strategies include the local linear model of LEAFAGE (see section 3.2), LIME and a baseline model. The baseline model always predicts the class predicted by f, on the instance being explained.

4.2 Datasets and black-box models

The different black-box ML models used are namely *Logistic Regression* (LR), *Support Vector Machine* with linear kernel (SVM), *Linear Discriminant Analysis* (LDA), *Random Forest* (RF), *Decision Tree* (DT) and *K Nearest Neighbour* with $K = 1$ (KNN). *Scikit-learn 0.19.2* [1] library was used to build these models with its default hyper-parameter unless stated otherwise.

These classifiers are applied to 5 different datasets with different number of features, rows and complexities (see table 1) and described below. Multi-class datasets with n classes are converted into n binary datasets of one-vs-rest fashion. We call a combination of a binary dataset and a classifier *a setting*. In total there are 54 settings.

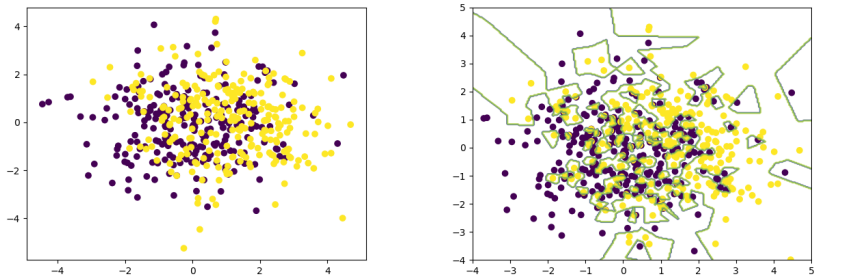
Dataset	Number of features	Number of rows	Number of classes	Amount per class
Iris	4	150	3	50/50/50
Wine	13	178	3	59/71/48
Breast Cancer	32	569	2	212/357
Bank Note	4	1372	2	762/610
AD	2	500	2	250/250

Table 1: UCI datasets used for quantitative evaluation

- *Iris*: This is a famous classification task of predicting the type of iris flower given it's different attributes about the petal and sepal.

[1] http://scikit-learn.org/stable/

- *Wine*: This dataset contains the chemical analysis (independent variables) of three different types of Italian wines (dependent variable).
- *Breast Cancer*: It is a binary classification task of classifying whether a breast mass is benign or malignant given different features about it.
- *Banknote*: This a binary classification problem of predicting the authenticity of a banknote as *fake* or *real*, given its different numerical characteristics.
- *Artificial dataset (AD)*: At last, we include a two-dimensional binary artificial dataset with highly non-separable classes (figure 9a). The instances of each class are sampled from two bi-variate normal distribution with different means ($[0, 0]$ and $[0, 1]$, respectively) and the same covariance matrix ($\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$). The KNN classifier trained on this dataset is visualized on figure 9b. The decision boundary is complex and highly non-linear.



(a) A highly non-linearly separable binary dataset.

(b) A KNN model trained on the *AD* dataset.

Figure 9

4.3 Results

Per setting, the average fidelity score along with the standard deviation in brackets is presented on table 2. The local linear model of LEAFAGE is computed with $i_{small} = 10$. Further, statistical tests per setting between the strategy with the highest mean and the rest of the strategies are performed. We perform a Wilcoxon signed-ranked test which tests the null hypothesis that two related paired samples come from the same distribution. The statistical tests are performed with a critical value of 0.05 with Bonferroni correction.

First, we view the performance of LIME and LEAFAGE versus the baseline model. Both methods have consistently average score higher than the baseline model over all settings.

Further, we establish whether there is a difference in performance between ML models with linear (SVM, LDA, LR) and non-linear (DT, RF, KNN) decision boundaries. We expect that the strategies perform better on the former models in comparison to the latter, because both LEAFAGE and LIME are based on linear models. Indeed, both strategies do perform better on the former models. This is very clear with the highly linearly non-separable *AD* dataset. On the models with linear decision boundaries these strategies have an average fidelity score of greater than 98%, while on non-linear models the scores are in the sixties.

Next, we look at the differences between LIME and LEAFAGE. Overall none of the strategies are consistently better than the other. On the models with linear decision boundary LIME scores significantly better than the LEAFAGE in 13 out of 27 setting. While on the non-linear models LEAFAGE strategies score significantly better than LIME on 7 out of 27 settings. The better performance of LIME on models with

Classifier Name	Strategy	Iris			Wine			BreastCa.	BankNote	AD
		Setosa vs rest	Versicolor vs rest	Virginica vs rest	Class 0 vs rest	Class 1 vs rest	Class 2 vs rest	Benign vs Malignant	0 vs 1	0 vs 1
LDA	Baseline	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)
	LIME	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	99.5 (1.0)	**100 (0.0)**	**100 (0.0)**
	LEAFAGE	**100 (0.0)**	96.1 (8.2)	**100 (0.0)**	100 (0.0)	96.0 (9.4)	100 (0.0)	**99.9 (0.3)**	99.9 (1.7)	98.6 (4.0)
LR	Baseline	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)
	LIME	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	**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)**	96.9 (10.6)	**100 (0.0)**	**97.1 (14.2)**	**100 (0.0)**	98.6 (7.8)	99.8 (0.9)	98.6 (4.0)
SVM	Baseline	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)
	LIME	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	**99.9 (0.6)**	**100 (0.0)**	**100 (0.0)**
	LEAFAGE	**100 (0.0)**	95.4 (9.9)	96.9 (10.6)	**100 (0.0)**	**100 (0.0)**	**100 (0.0)**	98.6 (7.8)	99.8 (0.9)	99.4 (1.2)
DT	Baseline	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)
	LIME	**100 (0.0)**	**96.0 (13.1)**	**100 (0.0)**	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	**100 (0.0)**	**81.5 (36.6)**	97.1 (10.6)	**92.9 (16.0)**	**85.8 (24.1)**	**100 (0.0)**	**86.5 (18.7)**	**98.7 (4.2)**	**65.0 (33.0)**
RF	Baseline	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)
	LIME	**100 (0.0)**	**97.1 (9.5)**	99.5 (1.1)	**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)**	**86.1 (29.9)**	**100 (0.0)**	**100 (0.0)**	**99.2 (3.7)**	**100 (0.0)**	**99.9 (0.8)**	**98.7 (3.8)**	**67.4 (32.9)**
KNN	Baseline	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)	50.0 (0.0)
	LIME	**100 (0.0)**	**98.0 (10.0)**	97.7 (8.3)	**98.0 (13.6)**	**62.8 (37.1)**	**60.5 (35.7)**	95.8 (8.2)	**100 (0.0)**	**65.6 (34.3)**
	LEAFAGE	**100 (0.0)**	**87.9 (29.2)**	**100 (0.0)**	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 2: The average local fidelity per setting with the standard deviation in brackets. The strategy with the highest mean along with other strategies that are not significantly different are denoted with a dark font color.

linear decision boundary could be because 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. This might explain the better performance of LEAFAGE over LIME on non-linear models.

In conclusion, LEAFAGE and LIME perform consistently better than the baseline model. Overall LIME performs better than LEAFAGE, on model with linear decision boundaries. However, LEAFAGE performs overall better than LIME on non-linear models.

5 Empirical Evaluation

To assess the real usefulness of LEAFAGE we perform a user-study. This user-study researches how useful the different parts of a LEAFAGE explanation are to the user. More specifically we evaluate the usefulness of example-based and feature importance-based explanations in terms of perceived aid in decision-making, measured transparency and persuasion.

Each participant is introduced to the user-study as follows:

> Imagine you are looking to buy a property. You are searching online and you find a couple of houses you like. Because it is a big investment, you are interested in the real value of the house. You find a smart application called LEAFAGE that estimates the value of a house automatically given its different features. For simplicity, let's say it predicts whether a property has low or high value. LEAFAGE provides four types of explanation, indicating why the house was predicted with a certain value. In this study we want to understand which type of these explanations are useful to the user in decision making.

5.1 Dataset used

The user-study is applied on the IOWA housing dataset [5]. It is an online free available public dataset that describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. It has been made available by the Ames City Assessor's Office in 2011. Different important characteristics along with the sale price per property are specified in the dataset.

We consider houses with sale price lower than 150.000$ and higher than 200.000$ as value *low* and *high*, respectively. Houses in-between those prices are removed, such that an ML model can be built that can predict the sale value (low or high) with high performance. A bad reasoning of the underlying ML model can affect the perception of an explanation negatively. To mitigate this influence, we choose a high performing ML model. The resulting dataset contains 619 houses with value low and 427 houses with value high. Furthermore, we use 5 features of each of these houses to estimate their value. These features are *Living Area*, *Year Built*, *Overall Quality*, *Bathroom Amount* and *Bedroom Amount*. They are chosen because they are understandable to the general public.

The dataset is split into two disjoint sets of training and testing set, which contain 70% and 30% of data, respectively. A SVM model with RBF kernel is built on the training set to predict sale value (low or high) of a house given its five features. This model has an AUC score of 98% on the testing set. The average local fidelity measure on the testing-set of LEAFAGE is equal to 98% with a standard deviation of 0.02%.

5.2 Experimental Design

The objective of the experiment is to investigate the effect of example-based and feature importance-based explanations extracted from LEAFAGE on the perceived aid in decision-making, measured transparency and persuasion.

5.2.1 Independent variable

We investigate 4 types of explanations namely *feature importance-based*, *example-based*, a combination of *example and feature importance-based* and finally providing no explanation as a baseline. An example of each explanation can be found in figure 10. Figure 10a shows an example of a house. Figures 10b, 10c, 10d, 10e show four types of explanations for the prediction of the house.

5.2.2 Dependent variables

First, we look into how the explanations are perceived by the participants in terms of aid in decision-making. We hypothesize that providing explanations behind predictions, aid more in decision making than providing no explanation. Example-based reasoning is known to be used in problem-solving e.g. diagnosing a patient [11]. But given LEAFAGE example-based explanation, it might not be easy to understand which features were important for the prediction. We further hypothesize that providing feature importance-based explanation in addition, will aid more in decision making. We propose four different variables that are important in an explanation such that the user can make an informed decision. The descriptions of these variables along with the null hypothesis are listed below.

1. Transparency: Whether the user understands the reasons behind a prediction
   $H_0 1$*: The median transparency score is the same for all explanation methods.*
2. Information sufficiency: Whether the user has enough information to make an informed decision
   $H_0 2$*: The median information sufficiency score is the same for all explanation methods.*
3. Competence: Whether the explanation corresponds to the user's own decision making logic
   $H_0 3$*: The median competence score is the same for all explanation methods.*
4. Confidence: Whether the explanation makes the user more confident about his/her decision
   $H_0 4$*: The median confidence score is the same for all explanation methods.*

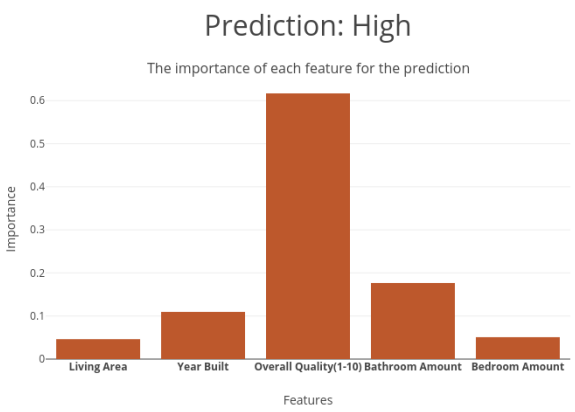
The previous hypotheses are about how the user perceives an explanation. Second, we would like to measure objectively how much knowledge the user gained about the underlying black-box ML model after looking at an explanation. We establish objectively whether the different explanation types have different effect on the insights that the users get about the underlying black-box model:
$H_0 5$*: The median measured transparency score is the same for all explanation methods.*

Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
192 m^2 (2076 ft^2)	2006	10	3	2

(a) House being explained

Prediction: High

(b) Explanation type: No explanation



(c) Explanation type: Feature importance-based

Prediction: High

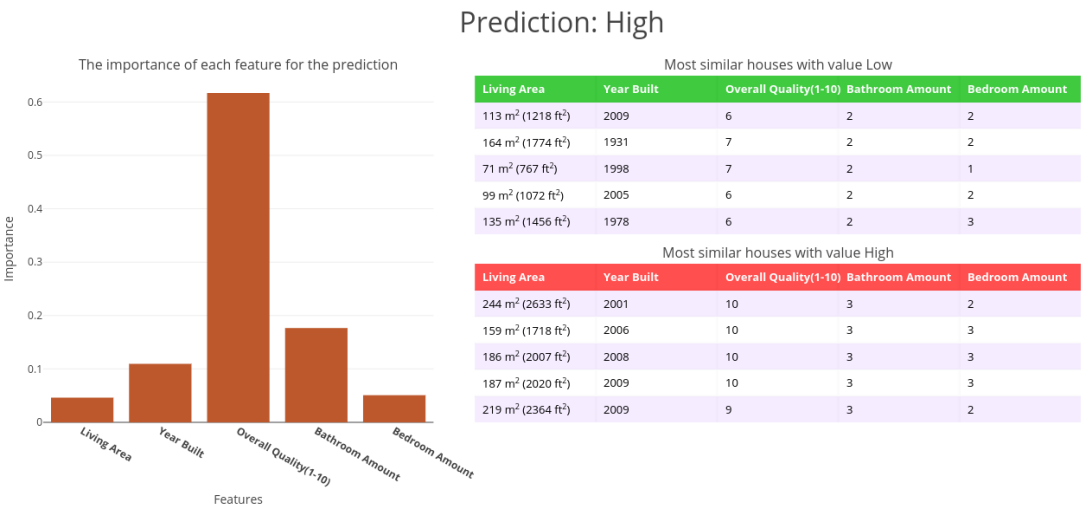
Most similar houses with value Low

Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
113 m^2 (1218 ft^2)	2009	6	2	2
164 m^2 (1774 ft^2)	1931	7	2	2
71 m^2 (767 ft^2)	1998	7	2	1
99 m^2 (1072 ft^2)	2005	6	2	2
135 m^2 (1456 ft^2)	1978	6	2	3

Most similar houses with value High

Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
244 m^2 (2633 ft^2)	2001	10	3	2
159 m^2 (1718 ft^2)	2006	10	3	3
186 m^2 (2007 ft^2)	2008	10	3	3
187 m^2 (2020 ft^2)	2009	10	3	3
219 m^2 (2364 ft^2)	2009	9	3	2

(d) Explanation type: Example-based



Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
113 m^2 (1218 ft^2)	2009	6	2	2
164 m^2 (1774 ft^2)	1931	7	2	2
71 m^2 (767 ft^2)	1998	7	2	1
99 m^2 (1072 ft^2)	2005	6	2	2
135 m^2 (1456 ft^2)	1978	6	2	3

Living Area	Year Built	Overall Quality(1-10)	Bathroom Amount	Bedroom Amount
244 m^2 (2633 ft^2)	2001	10	3	2
159 m^2 (1718 ft^2)	2006	10	3	3
186 m^2 (2007 ft^2)	2008	10	3	3
187 m^2 (2020 ft^2)	2009	10	3	3
219 m^2 (2364 ft^2)	2009	9	3	2

(e) Explanation type: Example and feature importance-based.

Figure 10: The first figure on the left shows the house being explained. The rest of the figures are the different types of explanations considered in the user-study.

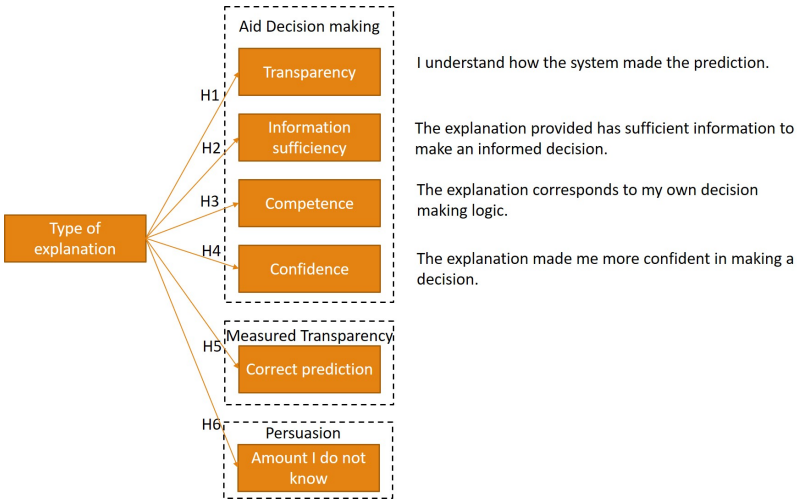


Figure 11: The dependent and independent variables.

At last, we establish objectively in what extent the participants are confident that they understand the the reasons behind a prediction, given each explanation type. Example-based reasoning is heavily used in law for the goal of justifying arguments, positions and decisions [11]. We hypothesize that LEAFAGE example-based explanation is more persuasive than feature importance-based explanation. Further, we hypothesize that a *example and feature importance-based* explanation is more persuasive than each separate.
$H_0$6*: The median persuasion score is the same for all explanation methods.*

An overview of all hypotheses can be found in figure 11. After looking at an house and an explanation for the prediction, the participants are asked to do two things. First, they have to rate the given explanation in Likert scale [19] according to transparency, information sufficiency, competence and confidence. The chosen scale has five values namely strongly disagree, disagree, undecided, agree and strongly agree (1-5 in the analysis). Next, the transparency is objectively measured by testing the participant. The participant is shown another house that is similar to the house being explained. He/she has to indicate what the system would predict as the sale value of this new house. The participant has the options of *low*, *high* or *I do not know*. The amount of correct answers per participant is used to compare the measured transparency between different explanations. If the participants are not sure about their answer they are recommended to choose *I do not know*. At last, persuasion is measured by counting the amount of *I do not know* answers per user. This shows whether the participants are confident enough in their ability to generalize the logic of the model to other houses, after seeing the explanation.

5.2.3 Participants

The participants for this user-study were recruited from Amazon Mechanical Turk (MTurk). In total 114 participants completed the survey. We perform the analysis of 86 participants who passed the attention checks that were present. The target group for this study is the general public. The demographics of the participants can be found in table 3. In terms of *Sex*, *Age* and *Highest level of education* the distributions seem well spread. Regarding *Region*, most of the participants are from the Americas. This could bias the results towards the preferences of the people of that region. Most of the participants are native or fluent English speakers. At last, 86% percentage of the participants have looked into buying a house.

5.2.4 Procedure

The different steps that the participants followed, are described in figure 12. Participants are first provided with a general introduction about the survey and asked to accept the terms of an informed consent. Next,

Sex		**Highest level of education**	
Male	45.3%	Less than high school	0.0%
Female	54.7%	High school	8.1%
Age		Some college (no degree)	15.1%
18 to 24	17.4%	Associate degree	15.1%
25 to 34	39.5%	Bachelor degree	48.8%
35 to 44	32.6%	Graduate degree	12.8%
45 to 54	17.4%	**Looked into buying a house**	
55 to 64	7.0%	Yes	86.0%
65 to 74	2.3%	No	14.0%
75+	0.0%	**Level of English**	
Region		Native/Fluent	89.5%
Africa	0.0%	Good	14.0%
Americas	80.2%	Satisfactory	0.0%
Asia	16.3%	Not very good	0.0%
Europe	3.5%	Bad	0.0%
Oceania	0.0%		

Table 3: Demographics information of the participants

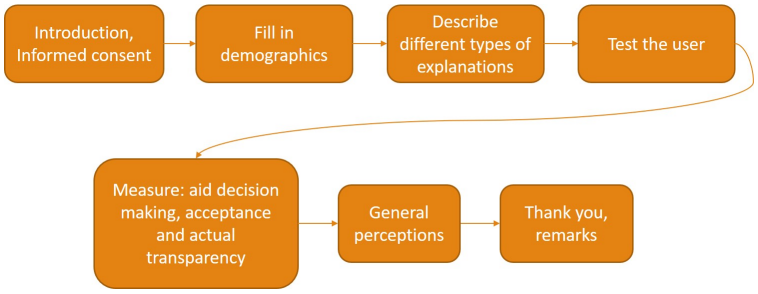


Figure 12: The procedure of the user-study.

they are provided with a demographic questionnaire. Further, the different types of explanations considered in this study are explained. Participants are then asked basic questions about the different explanation methods. The dependent variables described in figure 11 are measured next, on 40 explanations. Each participant sees 10 explanations per explanation type. The 40 houses being explained, were randomly chosen from the testing set. Moreover, all participants saw the same explanations in a randomized order. Before, the thank you and general remarks, the users are asked to provide their general perceptions for each type of explanation.

5.3 Results

5.3.1 Descriptive statistics

Table 4 shows an overview of the means, including the standard deviation in brackets, of all of the dependent variables. The first four columns are the scores (1-5) in Likert scale given by the participants on the dependent variables related to aid in decision making. The mean score of all explanation types are higher than providing no explanation over all columns. Furthermore, example-based and the combination explanation have a higher mean score than the rest.

The last two columns of table 4 are the means of the total score and the amount of 'I do not know' answers out of ten questions of each participants. The participants were most uncertain about their answer and have the least amount of correct estimation after seeing a feature importance-based explanation. Furthermore,

	Transparency	Info. Suff.	Competence	Confidence	Measu. Trans.	Persuasion
No Explanation	3.66 (1.03)	3.43 (1.17)	3.70 (0.96)	3.52 (1.1)	8.40 (1.48)	0.57 (0.96)
Feature importance	3.92 (0.85)	3.76 (0.97)	3.78 (0.91)	3.68 (1.04)	7.20 (1.66)	0.77 (1.21)
Example-based	4.07 (0.76)	4.02 (0.84)	3.96 (0.86)	3.98 (0.86)	8.83 (1.40)	0.15 (0.45)
Exam. and Fea.	4.13 (0.8)	4.10 (0.83)	3.93 (0.93)	3.98 (0.9)	8.56 (1.68)	0.27 (0.56)

Table 4: Means with standard deviation in brackets of the different dependent variables per explanation type.

	Dependent Variable	H statistic	p-value
$H_0 1$	Transparency	124.22	< 0.001
$H_0 2$	Information Sufficiency	202.71	< 0.001
$H_0 3$	Competence	54.83	< 0.001
$H_0 4$	Confidence	125.24	< 0.001
$H_0 5$	Measured Transparency	51.88	< 0.001
$H_0 6$	Persuasion	23.80	< 0.001

Table 5: Kruskal-Wallis H-tests performed on the dependent variables

example-based explanations lead to the most amount of correct predictions and the least amount of *I do not know*.

5.3.2 Hypothesis testing

All of the hypotheses are tested with a significance level of $\alpha = 0.05$ with Bonferroni correction.

Six Kruskal-Wallis H-tests [12] are performed on each of the dependent variables. The results are visualized in table 5. We reject $H_0 1$, $H_0 2$, $H_0 3$, $H_0 4$, $H_0 5$ and $H_0 6$ with p-value < 0.001.

Furthermore, a Dunn's post-hoc test [7] is performed to look at significant differences ($\alpha = 0.0083$[2]) in performance between pairs of explanation types. In terms of transparency, information sufficiency, competence and confidence both example-based and combination explanation perform significantly better than providing no explanation and feature importance-based explanation. No significant differences were found in performance between example-based and combination explanation regarding all four dependent variables. Feature importance-based explanation performs significantly better than providing no explanation in transparency, information sufficiency and confidence. But in terms of competence, feature-based does not perform significantly better than providing no explanation.

Moreover, in terms of measured transparency feature importance-based explanation performs significantly worse than of the rest of the explanation types. The amount of correct estimation was statistically similar between pairs of providing no explanation, example-based and combination explanation.

At last, regarding persuasion example-based explanation performs significantly better than providing no explanation and feature importance-based explanation. The combination explanation performs significantly better than feature importance-based explanation but no significant different was found in comparison to example-based explanation and providing no explanation. Finally, the persuasion score was similar between feature importance-based explanation and providing no explanation.

[2]Applying Bonferroni correction: 6 comparison per dependent variable leads to a significance level of $0.05/6 = 0.0083$

5.3.3 General remarks about the explanation types

After the experiment the participants were asked to state their likes and dislikes for each of the explanation types.

Regarding getting no explanation, the participants liked that only getting the prediction was straightforward, easy to understand, simple, to the point and allowed them to make quick decisions. To quote a few: "Easy to digest and comprehend and conveys the key information that one wants to obtain", "I do not have to make any analysis", "No need to overthink" and "Easy to make a decision quickly". Furthermore, the participants disliked the lack of information supporting the prediction, the fact that the prediction provides no context and its vagueness. Here are a few quotes expressing those concerns: "Need complete trust in the system to find it helpful", "Could be seen as just a guess with no rational behind, doesn't seem as legitimate" and "I don't want a simple rating without anything to back it up. Show me the numbers and let me decide what is important".

The participants found that feature importance explanation type was straightforward in showing how the sale value was determined and easy to read. They liked the visual aspect of the graph. The following quote summarize their likes of the explanation type: "I like the easy to grasp and digest nature of this visual depiction of the rationale behind the rating used to evaluate a house." However the participants also had some concerns about this explanation type. They found this explanation type not detailed enough to perform well on the *measured transparency* part of the study. It was not clear to them how the importance really related to the prediction of the house. Moreover, it was hard for them to understand the threshold of a feature which changes the prediction of a house. One participant wrote "If I had to list a dislike, it would be that there is no explanation WHY these features are of importance to the prediction model? How are they rated compared to one another? Do some features hold more weight overall? Because it seemed like it at times."

Regarding example-based explanation the participants liked the fact that they could compare between similar houses with different sale values. They could see how the feature values differ between houses with different sale values. However, the participants both liked and disliked this explanation type because of the amount of information present in the tables. The large amount of raw data helped them to understand in detail why the prediction was made but lot of numbers were harder to digest and made the information look cluttered. Moreover, they found it hard to figure out which features were important. A participant summarized the pros and cons as follows: "I like this explanation because it provides references as to what would be considered a high or low valued home. With this layout I can easily compare the target home to similar homes on the market and decipher what features are associated with the high or low value homes. The only downside this explanation is that he does not provide a detailed explanation as to the importance of each feature in making the decision."

The participants liked the combination of example-based and feature importance-based explanation but also stated that the amount of information can be overwhelming. Some participants said that they only looked at one chart and ignored the other. The overall remarks of this explanation type is nicely summarized by one participant as follows: "I like this explanation type because it incorporates both an easy to digest visual depiction i.e. the bar graph and the raw data used to build this algorithm calculator. This juxtaposition of key elements of the algorithm facilitates the ability of users to obtain a more detailed and informed idea regarding the backbone of the algorithm. I guess one potential downside is that low information users may be turned off and/or intimidated by the juxtaposition of a graph and data chart."

At last, the participants were asked to give general remarks about the survey. The participants stated that even though the survey was long they could stay engaged because of the *measured transparency* part. One participant wrote: "The survey was lengthy, but engaging and fun to participate in. I strongly believe that information like this will be a great benefit to home buyers."

6 Conclusion

In this study we developed a new method called LEAFAGE that provides contrastive example and feature-based explanations for the predictions made by a black-box ML model. Furthermore, for the evaluation of LEAFAGE two evaluation methods were used. The first method evaluated whether LEAFAGE explanations reflect the true reasoning of the underlying black-box ML model (*local fidelity*). Second, we looked into the usefulness of the explanations from the user's point of view, through conducting a user-study.

We proposed a new method to evaluate the local fidelity of a local interpretable model that mimics the black-box classifier, in the neighbourhood of the instance being explained. This evaluation method was used to compare the local linear model of LEAFAGE with the current state-of-art method LIME [23]. The evaluation was done for 5 datasets with different characteristics and 6 different ML models. The results showed that overall LIME performed better than LEAFAGE, on models with linear decision boundaries. On the other hand, LEAFAGE performed overall better than LIME on non-linear models. This could be due to the fact that LEAFAGE takes the local decision boundary into account in the calculation of the neighbourhood.

By performing a user-study, we evaluated the usefulness of example-based and feature importance-based explanations extracted from LEAFAGE, in terms of the perceived aid in decision making, acceptance and measured transparency. The context of the user-study was to help the participants in estimating the value of a house, by providing a prediction made by a ML model and an explanation. Example-based explanation performed significantly better than providing no explanation and feature importance-based explanation, regarding perceived transparency, information sufficiency, competence and confidence. This was expected because example-based reasoning is regarded as a natural way of solving problems in our daily lives [3, 1]. Moreover, we had expected that adding information about the importance of features would increase the perceived aid in decision making, but no significant difference was found. This could be because the amount of information became overwhelming and seemed cluttered as the comments of the participants suggested. Feature importance-based explanation performed significantly better than providing no explanation in terms of transparency, information sufficiency and confidence. However no significant difference with providing no explanation was detected regarding competence. This result suggests that feature importance-based explanation does not align more with the decision making of the participants than providing no explanation.

Second, we expected that example-based explanations would lead to higher measured transparency than providing no explanation but no significant difference was found. Moreover, feature importance-based explanation led to significantly less amount of correct estimation, compared to the rest of the explanation types. These results suggest that there is a discrepancy between the perceived transparency and the measured transparency, because the participants preferred both feature-importance and example-based explanation over providing no explanation. This could be because the task was too easy and people were able to guess the sale value of the house even without seeing an explanation. The fact that the participants performed worse after seeing a feature importance-based explanation than providing no explanation, suggests that feature importance-based explanations confuse the users in how they can generalize the provided classification logic in the explanation to other instances. In the comments some participants noted that it was not clear how the importance of a feature really relates to the prediction of the house.

At last, as expected, example-based explanations persuaded the users significantly more, in believing that they understood the prediction, than providing no explanations or feature importance-based explanation. Adding feature importance-based explanation to example-based explanation did not lead to significantly higher score for persuasion. Moreover, the persuasion score was similar between feature importance-based explanation and providing no explanation.

In conclusion, LEAFAGE explanation performed overall better than the current state-of-the-art method LIME on non-linear models, in terms of local fidelity. The empirical evaluation showed that overall the participants perceived receiving explanations behind a prediction as more helpful than providing no explanation for the goal of decision making. However when the 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, as future work a more comprehensive test could be used to measure the actual transparency. Moreover, the participants scored significantly lower on feature importance-based explanation compared to providing no explanation. This is an important result, which suggests that feature importance-based explanation confuses the user more about the prediction reasons than providing no explanation. Further, example-based explanation significantly performed 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. At last the results showed that example-based explanations persuaded the users significantly more, in believing that they understood the prediction, compared to feature importance-based explanation.

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 Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(4):421–434, 2011.

[3] Isabelle Bichindaritz and Cindy Marling. Case-based reasoning in the health sciences: What's next? *Artificial intelligence in medicine*, 36(2):127–135, 2006.

[4] 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.

[5] Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011.

[6] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

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

[8] 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.

[9] P.H.N. Gill. *Introduction to Machine Learning Interpretability*. O'Reilly Media, Incorporated, 2018.

[10] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models. *arXiv preprint arXiv:1802.01933*, 2018.

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

[12] 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.

[13] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Comparison-based inverse classification for interpretability in machine learning. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 100–111. Springer, 2018.

[14] 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.

[15] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990.

[16] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[17] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

[18] 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.

[19] Geoff Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632, 2010.

[20] Pearl Pu and Li Chen. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20(6):542–556, 2007.

[21] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 157–164. ACM, 2011.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[24] Michael M Richter and Rosina O Weber. *Case-based reasoning*. Springer, 2016.

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

[26] 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 *International conference on machine learning*, pages 2048–2057, 2015.