

Characterizing Data Scientists' Mental Models of Local Feature Importance

Citation for published version (APA):

Collaris, D., Weerts, H. J. P., Miedema, D., Wijk, J. J. V., & Pechenizkiy, M. (2022). Characterizing Data Scientists' Mental Models of Local Feature Importance. In *NordiCHI '22: Nordic Human-Computer Interaction Conference Article 9* Association for Computing Machinery, Inc. <https://doi.org/10.1145/3546155.3546670>

DOI:

[10.1145/3546155.3546670](https://doi.org/10.1145/3546155.3546670)

Document status and date:

Published: 08/10/2022

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Characterizing Data Scientists' Mental Models of Local Feature Importance

Dennis Collaris*
Eindhoven University of Technology
The Netherlands
d.a.c.collaris@tue.nl

Hilde Weerts*
Eindhoven University of Technology
The Netherlands
h.j.p.weerts@tue.nl

Daphne Miedema
Eindhoven University of Technology
The Netherlands
d.e.miedema@tue.nl

Jarke J. van Wijk
Eindhoven University of Technology
The Netherlands
j.j.v.wijk@tue.nl

Mykola Pechenizkiy
Eindhoven University of Technology
The Netherlands
m.pechenizkiy@tue.nl

ABSTRACT

Feature importance is an approach that helps to explain machine learning model predictions. It works through assigning importance scores to input features of a particular model. Different techniques exist to derive these scores, with widely varying underlying assumptions of what importance means. Little research has been done to verify whether these assumptions match the expectations of the target user, which is imperative to ensure that feature importance values are not misinterpreted. In this work, we explore data scientists' mental models of (local) feature importance and compare these with the conceptual models of the techniques. We first identify several properties of local feature importance techniques that could potentially lead to misinterpretations. Subsequently, we explore the expectations data scientists have about local feature importance through an exploratory (qualitative and quantitative) survey of 34 data scientists in industry. We compare the identified expectations to the theory and assumptions behind the techniques and find that the two are not (always) in agreement.

CCS CONCEPTS

• **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; **Machine learning**; • **Human-centered computing**;

KEYWORDS

Interpretability, Explainable AI, Feature importance

ACM Reference Format:

Dennis Collaris, Hilde Weerts, Daphne Miedema, Jarke J. van Wijk, and Mykola Pechenizkiy. 2022. Characterizing Data Scientists' Mental Models of Local Feature Importance. In *Nordic Human-Computer Interaction Conference (NordiCHI '22)*, October 8–12, 2022, Aarhus, Denmark. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3546155.3546670>

*Both authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution International 4.0 License.

NordiCHI '22, October 8–12, 2022, Aarhus, Denmark
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9699-8/22/10.
<https://doi.org/10.1145/3546155.3546670>

1 INTRODUCTION

As machine learning is increasingly used for high-stakes decision making, it is essential the models we built can be held up to scrutiny. To this end, the field of eXplainable AI (XAI) has introduced various techniques that aim to support understanding the decisions of complex machine learning models. Many of these techniques fall under the umbrella of 'feature importance': techniques that calculate a scalar value for each feature to provide insight into the importance of that feature towards either the overall behavior of the machine learning model (i.e., global) or an individual prediction (i.e., local). In this work, we focus specifically on local feature importance.

Local feature importance techniques have seen widespread adoption, popularity, and success in solving real world problems [8, 26]. However, a systematic way to evaluate and compare these methods remains elusive, as the qualities of an adequate explanation of a machine learning model are inherently subjective. Different feature importance techniques make different assumptions about the properties that a good explanation should have, which can cause them to be inconsistent or even contradictory [10]. In addition, recent work has critiqued many of the existing techniques on various accounts: being misleading [12], lacking robustness [2], and not enabling action [23].

In this work, we explore the extent to which experts' assumptions about local feature importance match existing techniques. Specifically, our main contributions are: 1) a comprehensive overview of important properties affecting the interpretation of feature importance; 2) a qualitative characterization of how data scientists in industry define feature importance; 3) a quantitative survey exploring the expectations of experts of the identified properties; and 4) a set of recommendations for XAI researchers to better match feature importance techniques with expectations of data scientists. While we found most identified properties are expected, some expectations were conflicting, or varied a lot amongst participants. This warrants careful consideration.

The remainder of this paper is structured as follows. We first describe related work (Section 2 and 3). In Section 4, we lay out several potentially misleading properties of local feature importance techniques. Section 5 details our survey design. Section 6 and Section 7 cover the results, followed by a discussion (Section 8) and concluding remarks (Section 9).

2 LOCAL FEATURE IMPORTANCE

We define feature importance as any quantitative assignment of importance or influence to the features used by a machine learning model. There are two primary ways in which feature importance is construed. *Global* feature importance techniques attribute importance to a feature in relation to the model or its predictions as a whole. *Local* feature importance techniques, on the other hand, produce explanations that pertain to the prediction of a single data point. The focus of this paper is primarily on local feature importance. We distinguish between two types of techniques to compute these values: gradient-based and ablation-based feature importance.

Gradient-based. This type of feature importance techniques assume that features are important when small changes in feature value result in a (relatively) big change in model prediction. As such, gradient-based feature importance values can be interpreted in a similar fashion as the coefficients in linear regression models, which are widely considered to be interpretable (global) explanations [13, 35]. We discuss three examples of influential gradient-based techniques.

Baehrens et al. [3] show that an exact derivative (i.e., gradient) of a model can be used as feature importance. As an exact derivative may not always exist, they use a Parzen window surrogate model to mimic the reference model, and use the derivative of that surrogate model to generate feature importance vectors.

Next, LIME [30] is a very popular technique that approximates the gradient by training a local interpretable surrogate model on generated samples, weighted by the inverse distance to the instance to be explained. If a linear regression surrogate model is used, the coefficients of that model approximate the derivative of the model.

Finally, saliency maps are a gradient-based explanation technique specifically targeted to neural networks trained on image data. These techniques aim to show which pixels in the input image were most relevant for the prediction, by computing the gradient of the neural network directly using back propagation (e.g., Grad-CAM [32]).

Ablation-based. These techniques assign feature importance by comparing model predictions when a feature value is present to when it is absent (i.e., *ablation*). It is generally not possible to simply remove a feature value from an existing model without changing the model’s parameters. As such, ablation-based feature importance approaches require a method to simulate absence of a feature value.

Shapley-value based approaches [21, 25, 33] pose the distribution of feature importance as a cooperative game, where each feature value is a player. In order to capture the influence of interactions between features, Shapley-value based approaches consider how the model prediction changes for each subset, or ‘coalition’, in the power set of features. Next, Shapley-value based approaches compute the change in prediction, or ‘value’ of each subset by averaging across all possible feature values of the features that are not part of the subset under consideration.

Zeiler and Fergus [36] present another ablation-based importance technique specifically for image classification. Here, importance is computed based on the extent to which iteratively masking input pixels with a gray value changes the prediction output.

Gradient-based and ablation-based approaches seem similar: both ascribe importance of a feature value based on how the model’s

prediction changes when the feature value changes. However, where gradient-based approaches consider the *rate* of change in the output, ablation-based techniques consider the *magnitude* of the change. Consider the example shown in Figure 1, which shows the predicted outcome $f(x)$ given a value x for a feature. The base rate denotes the average outcome over all values of x . For $x = 3.5$, we see a large difference of $f(x)$ with the base rate, but a small gradient (i.e., slope); for $x = 6$ this pattern is reversed. Here we see that a small perturbation changes the prediction significantly, whereas removal of the feature has almost no effect. As we describe in Section 4, this difference results in a very different interpretation of feature importance.

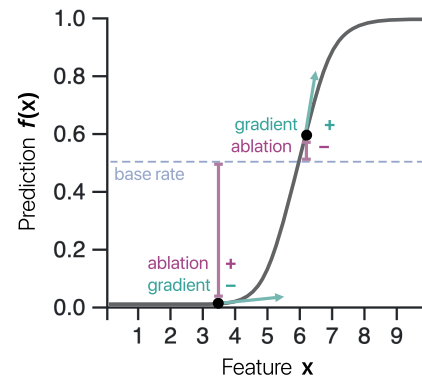


Figure 1: Gradient-based compared to ablation-based feature importance. Two points highlight for which the feature importance scores of both techniques vary widely.

3 RELATED WORK

Our study is closely related to recent work on mental models in XAI. Additionally, our exposition of potentially misleading properties is in line with recent critiques of local feature importance approaches.

3.1 Mental Models of Explanations

In the context of human-computer interaction, a *mental model* is a user’s belief about how the target system works [22]. Mental models are typically contrasted with a system’s *conceptual model*: a representation of the system as intended by the designer [29]. Erroneous mental models can lead to behavior with unintended consequences [29]. In XAI, explanations are often viewed as tools to increase the accuracy of a user’s mental model of the machine learning model [14, 22]. In the present paper, we are concerned with mental models of the explanation technique itself. If a user’s mental model of an explanation technique is inaccurate, this can result in misuse or misinterpretations. For example, in a recent evaluation of explanation tools, Kaur et al. [20] find that practitioners who have (partially) accurate models of an explanation tool make more careful decisions compared to those who take the visualizations at face value. We build upon these findings and set out to characterize a more detailed mental model of local feature importance.

3.2 Problems with Feature Importance Techniques

Several scholars have critically examined the underlying assumptions and intended interpretation of local feature importance techniques. A recurring topic of interest is the faithfulness of explanations: the extent to which the explanation approximates the prediction of the black-box model [31]. One of the main assumptions of LIME is that, even if a model is complex globally, it might be possible to estimate the model's gradients locally. However, even locally, it may not be possible to accurately estimate the gradients. Additionally, Lipton [24] remarks that for differentiable models (e.g., neural networks) the added value of LIME over raw gradients is unclear.

Another regularly discussed issue is the ambiguity of axiomatic 'desirable' properties of feature importance, which can lack proper justification [24] and contextualization [15]. For example, Kumar et al. [23] question the justification of the additivity constraint imposed by SHAP. Regarding contextualization, XAI techniques are typically developed without a specific use case in mind, even though the effectiveness or usability of a technique likely varies across scenarios [11, 13]. For example, several studies have shown a limited utility of using local feature importance for improving accuracy in decision-making by domain experts [19, 34], whereas other studies show that systems that rely on feature importance can lead to novel insights [16], faster decision-making [19], and effective feature selection [1].

When feature importance scores are misinterpreted, this could lead to wrong conclusions. For example, a data scientist may (wrongly) conclude that a high feature importance score retrieved from an ablation-based technique indicates that increasing the feature value will increase the model's score. In this paper, we examine to what extent several of such possibly misleading properties of local feature importance scores match with data scientists' expectations.

4 PROPERTIES OF LOCAL FEATURE IMPORTANCE

Based on our survey of related work, we provide an overview of several properties of local feature importance values that could influence how they can be (mis)interpreted. These properties will provide the basis for our survey.

As there are many ways in which feature importance can be characterized, our overview is inevitably non-exhaustive. The discussed properties were selected based on the extent to which they are properties of feature importance scores in absence of a specific application. For example, whereas faithfulness is an important property of explanations, the concept is ill-defined and cannot be assessed in absence of a specific machine learning model. Similarly, the usability of feature importance scores can only be assessed in relation to a specific task.

P1. Actionability. Feature importance computed through gradient-based approaches can be regarded as an approximation of the derivative of the model's predicted score over the feature. If the approximation is sufficiently accurate (locally), gradient-based feature

importance can be interpreted as (locally) *actionable*¹: if a feature is important, an action (i.e., a small change in feature value) will affect the model's score [20]. The same does not hold for ablation-based approaches. For example, a high SHAP value implies that, on average, the model's score would have been different if the instance would have had another feature value. However, presented as an average, it does not indicate *how* the feature value should have been different - there could have been one specific alternative feature value with a very different score or an entire range of feature values with varying scores.

P2. Causality. "Correlation does not imply causation." This expression is often used to warn data analysts of misinterpreting statistical correlations as causal relationships. Most machine learning models are statistical models. While (local) feature importance values may help to formulate new hypotheses, they should never be interpreted directly as causal relationships between features and the target variable. Clarification of the non-causal nature of feature importance is especially important when the explanations are used as decision-support for less experienced users.

P3. Stability. Several local feature importance techniques rely on randomly sampled feature value perturbations. Additionally, explanations may be sensitive to the choice of parameters. As such, the resulting feature importance values may differ across subsequent runs of the explanation algorithm [37]. We refer this as the *stability* of the explanation. Instability may cause users to be reluctant to use explanation methods [17].

Instability can be mitigated to some extent by increasing the number of samples [12], but this depends on the dimensionality of the data and affects running time. Regardless, for complex models, feature importance values will always constitute an approximation of the model's underlying prediction-generating mechanism. Consequently, there are usually various alternative (and equally valid) explanations for the same prediction [9]. However, feature importance is typically presented as a single value per feature, which may disguise the inherent uncertainties in how the values are derived [15].

P4. Robustness. The robustness of an explanation technique considers the similarity of explanations for similar instances [2]. This means that if feature values are perturbed slightly, the explanation is not changed unless the perturbations also strongly change the prediction. The property is closely related to stability. The main difference is that stability considers sensitivity to *parameters*, whereas robustness considers sensitivity to the *input*. Given their reliance on input perturbations for computing feature importance, it is perhaps unsurprising that if we rely on a relatively small number of perturbations, both SHAP and LIME can yield varying and inconsistent explanations for more complex models [2].

More generally, as the complexity of the model increases, it becomes more challenging to determine whether variations in feature importance should be attributed to the erratic behavior of the explanation method or the underlying machine learning model. Should

¹Note we use the definition of actionable as introduced by Kaur et al. [20]. In different contexts actionability may refer to other things, such as practical usefulness, or (in counterfactual explanations) the practical feasibility of changing a feature value for an individual.

we expect explanations to be robust at all? If the purpose of an explanation is to understand the underlying data, robustness may be desirable, as we are more interested in consistent patterns. However, if the purpose is model validation, ‘robust’ explanations may disguise unexpected model behavior.

P5. Selectivity. Research from social sciences shows that people do not expect explanations to provide a complete account of all causes for an event. Instead, people select a subset of causes for the explanation they believe to be the most important [28]. A feature importance method can be selective by limiting the explanation to the most important features (where the exact meaning of ‘important’ depends on the method). For example, common implementations of LIME use L1 regularization by default, to reduce the number of features in the explanation [30]. In contrast, SHAP explanations will include all features in the final explanation. Importantly, selectivity can be in tension with the faithfulness of an explanation: a very simple explanation is not able to fully capture the complexity of the model’s decision logic.

P6. Additivity. Lundberg and Lee [25] introduce the concept of additive feature attribution methods as a set of feature importance techniques that can be interpreted as a decomposition of the model’s predicted score over all features, resulting in one value per feature. Additivity can potentially lead to misleading interpretations. If features have a strong statistical relationship in the data set, it is unclear how feature importance of individual features should be interpreted. Similarly, features may strongly interact with each other in the machine learning model. For example, how should we distribute importance in an additive fashion if the model presumes an ‘XOR’ relationship between two features? As remarked by Hancox-Li and Kumar [15], users may interpret feature importance to represent solely univariate effects, which is fundamentally misleading when a non-additive model is explained. Again, we see that summarizing complex model behavior in a few numbers may be an oversimplification.

P7. Proportionality. We consider a feature importance technique to be proportional if the sum of feature importance values is proportional to the output of the original model. To illustrate, consider the relation between feature importance vectors from LIME and SHAP and the predicted score of the model:

$$\begin{aligned} \text{Gradient-based (e.g., LIME): } \hat{y} &= \alpha + \sum_i \beta_i X_i & \text{(a)} \\ \text{Ablation-based (e.g., SHAP): } \hat{y} &= \epsilon + \sum_i \phi_i & \text{(b)} \end{aligned} \quad (1)$$

Because the base rate ϵ of Shapley values is constant, the sum of Shapley feature importance values ϕ is directly proportional to the original model prediction score \hat{y} (with offset ϵ). However, for LIME, the feature importance values first need to be multiplied by the feature values, and then added to the intercept α that is different for each instance. As an (arguably counter-intuitive) consequence, features tend to have low importance when the model is very certain, and features tend to be more important when the model is very uncertain (notable in Figure 1).

P8. Sampling Distribution. Various feature importance techniques rely on perturbing feature values, which requires a predefined distribution of possible feature values. We can distinguish two types of sampling distributions [23] with a different underlying intuition that affects how the resulting feature importance scores can be interpreted.

Interventional distributions allow for sampling across all possible feature values for each feature independently, irrespective of whether the resulting combination of feature values is likely to occur in the data. Chen et al. [7] consider an interventional approach appropriate when the goal is to understand the model independently of the data, as a mathematical function that maps input to output. That is, if we view the model as a mathematical function that maps an input to output, computing importance based on out-of-distribution samples seems acceptable. However, when importance is used to identify relationships that hold true in the data, interventional distributions can be misleading [18]. In particular, this can result in out-of-distribution samples. In these cases, Hooker et al. [18] propose to sample feature values from distributions that are *conditional* on the remaining features. As a result, any perturbed instance is consistent with the original data distribution.

These different approaches towards sampling reveal fundamentally different views of what it is that feature importance explains. Conditional distributions consider the informativeness of the feature, given the structure of the training data, whereas interventional distributions quantify the sensitivity of the model to a feature, regardless of the underlying data.

5 METHODOLOGY

Our goal is to explore data scientists’ mental models of (local) feature importance values and their implications for existing techniques. To this end, we pose the following research questions:

- RQ1** How do data scientists define feature importance?
- RQ2** What are the expectations of data scientists with respect to properties of local feature importance?

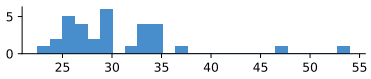
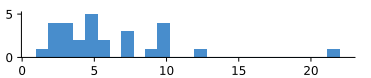
To answer our research questions, we conducted an online survey amongst data science professionals, using an exploratory mixed-methods survey approach.

5.1 Participants

We recruited participants using snowball and convenience sampling strategies. This means we invited industry acquaintances to participate in our survey and asked them to suggest and forward the survey to colleagues. The study was approved by our institution’s Ethical Review Board (ERB). Participants were not compensated for their contribution.

Each participant was presented with a consent form detailing the purpose and process of the study. After giving consent, participants were asked to supply basic demographic information (age, country of residence, gender identity) as well as their current job title and years of experience in the field of data science. 34 participants filled out the survey. Details about participant sample (demographics, role, experience in data science, and prior familiarity with feature importance techniques) are shown in Table 1.

Table 1: Summary of the participant demographics.

Type	Answers (count)
Gender	Male (20), Female (12), Prefer not to disclose (2), Prefer to self-describe (0)
Age	
Location	Netherlands (24), United States (2), Prefer not to disclose (2), Colombia (1), India (1), Singapore (1), Spain (1), Switzerland (1), United Kingdom (1)
Role	Data scientist (18), (Data science) researcher (6), Software/data/AI engineer (3), PhD candidate (3), (Data science) consultancy (2), AVP (Assistant Vice President) (1), Prefer not to disclose (1)
Data science experience (years)	
Familiarity	Linear regression coefficients (30), Random Forest feature importance (29), SHAP/Shapley values (23), LIME (23), Permutation importance (18), Saliency maps (e.g., GradCAM) (11), treeinterpreter (9), Anchors (2), DeepLift (1)

5.2 Survey

We set up an online survey that took between 20 and 30 minutes to complete. The survey started with the demographics questions as mentioned earlier. The remainder consisted of three parts (more detail in the supplemental material):

Feature Importance. To answer **RQ1**, participants were asked open questions to explain their interpretation of feature importance, (**Q1**) in the context of machine learning, (**Q2**) for a (trained) machine learning model, and (**Q3**) for an individual prediction. Additionally, participants were asked to express their opinion on the value of feature importance (**Q4**) and to describe a specific use case in which feature importance may support a process or workflow (**Q5**).

Expectations of Properties. To answer **RQ2**, participants were asked to indicate their expectations of local feature importance. We showed them five sets of statements (18 in total), each corresponding to the properties identified in Section 4. These statements were made more concrete through a running example about a medical model predicting risk of complications, based on (unspecified) medicine levels. Next, we asked participants to what extent they agreed with these statements, in the form of a 5-point Likert scale, ranging from ‘strongly disagree’ to ‘strongly agree’. Additionally, participants could provide optional textual comments to motivate and explain their answer.

The survey items were derived by applying each of the properties to a specific example. Face validity of the items was established by piloting the survey with two data scientists. To avoid primacy bias, we repeated each binary statement (e.g., either changes a little, or changed a lot) with reverse wording in the survey, and randomized the order of these two options. Repetition also enables us to verify the internal validity of the items, as we expect them to be opposing.

Familiarity. In the final part of the survey, participants were first asked to list all feature importance techniques they have used in their work. Once they had filled this in, they were provided with a list of specific techniques and asked to indicate which of these techniques they were familiar with.

5.3 Data Analysis

Qualitative Data Analysis. We performed a thematic analysis [4, 6] of the participants’ textual comments to identify which topics and aspects were reoccurring. The analysis consisted of an iterative qualitative coding process, characterized by alternate phases of coding, discussing and identifying (sub)-themes. Initially, the first three authors read and re-read the comments in order to identify potential themes. We used both inductive and deductive reasoning, the latter based on the identified properties of local feature importance methods. The second level of analysis involved reviewing the initial codes and identifying overarching elements. This process was repeated another two times, refining codes and themes.

Quantitative Data Analysis. We globally explored patterns by visualizing the results using divergent stacked bar charts, as shown throughout Section 7. Charts are annotated with p-values computed using the Mann–Whitney U test, chosen for its suitability to low frequency independent ordinal samples (e.g., Likert). We compare the distribution of answers against only-neutral answers, and report if the answers are skewed towards agreement or disagreement. We use significance level $p < 0.05$ (indicated with *) and correct for multiple comparisons using the Bonferroni method with $m = 18$, rejecting the null hypothesis at $p < \frac{0.05}{18}$ (indicated with **).

6 QUALITATIVE RESULTS

The thematic analysis of five questions yielded 10 themes, summarized in Table 2. Below, we discuss the definition of each theme, and the codes that belong to it. The first three themes concern the types of perspectives that our participants have about feature importance, giving us insights into their mental models. The latter themes (**T4–T10**) reflect the pros and cons of feature importance identified by our participants. One of our participants did not answer the open questions, leaving 33 responses for the qualitative analysis.

T1. Locality. The theme of locality focuses around the questions of whether our participants describe feature importance largely as a local technique, a global technique, or whether it concerns models in general. Most of the participants ascribed to a single perspective, five participants mentioned aspects of two different perspectives.

Participants who interpreted feature importance as mainly local (9 of 33), described feature importance from the perspective of a single prediction. For example, participant 17 mentioned “[It measures] how much a feature contributes to a prediction.” and participant 23 wrote: “[It] measures how much of the output prediction is explained by a given feature.”

Participants who think of feature importance as mostly global (20 of 33) discussed the technique as applied to ‘a model’. Examples of participant answers include “It’s a scoring model that scores the importance/impact of each feature on the outcomes of a ML model.” and “How much influence and to what degree every feature has in the decision process of the machine learning model.”

Table 2: Summary of identified themes, which questions the themes applied to, the prevalence of the themes, and a description.

Theme	Q1	Q2	Q3	Q4	Q5	Prev.	Description
T1. Locality	×					100%	FI explains a prediction, a model, or any model.
T2. Explanandum	×	×	×			87.9%	FI explains quality, informativeness, or predictions.
T3. Underlying mechanism	×	×	×			48.5%	FI is a gradient or ablation-based method.
T4. Understanding				×	×	72.7%	FI helps to understand a model or data.
T5. Feature selection				×	×	33.3%	FI enables feature selection.
T6. Debugging				×	×	42.4%	FI enables debugging.
T7. Trust and fairness				×	×	27.3%	FI enables identifying bias.
T8. Decision-making				×	×	24.2%	FI supports decision-making.
T9. Improve performance				×	×	12.2%	FI enables improving performance.
T10. Downsides				×	×	21.2%	Downsides of feature importance (w.r.t. properties in Section 4).

Finally, there is the even broader perspective of the importance of a feature towards any possible model for the data. In these answers (9 of 33), the words ‘prediction’ and ‘model’ are typically not present at all. For example, participant 13 writes: *“Feature importance is the contribution of each feature to modeling decisions across an adequate sample of data.”*

T2. Explanandum. This theme describes what our participants believe that a feature importance score captures. It contains two main categories. The first group of participants supposes that feature importance explains the quality of the model (12 out of 33 participants). The second group is of the opinion that feature importance explains the predictions of the model (27 out of 33 participants). These two groups are not mutually exclusive, some participants’ remarks contained elements of both categories (6 out of 33).

For the participants discussing feature importance as a measure of quality, terms we categorize to indicate model quality include informativeness, accuracy and predictive power. For example, answering Q2 on what global feature importance means to them, participant 10 mentions: *“That this feature in general is quite informative for the machine learning model.”* Regarding the influence and predictive power of a feature, participant 18 describes feature importance as *“[...] how influential a feature is within the context of a machine learning problem. In other words, if the feature were to not be used, how big is the impact on the predictive performance for machine learning models within the problem context?”*

The group that views feature importance as a reflection of a prediction uses words such as outcome, decision boundary, model input and output on top of the term prediction. For example, participant 25 writes that *“It means that the feature is important in terms of the data structure [...] that influences the outcome, or it is highly correlated with the outcome.”* and participant 27 writes *“Feature importance expresses the importance of a feature in the relation between input and output.”*

T3. Underlying mechanism. This theme describes which of the two identified mechanisms in Section 2 the users’ description matches. Only a part of participants’ answers (16 of 33) clearly indicated properties related to this theme.

A *gradient-based* perspective was indicated by mentioning permutations or small changes to the input data. This perspective was held by 8 of our 33 participants. Participant 30 describes feature importance as *“Perturbing the value of this feature just a little bit,*

heavily influences the outcome.” and participant 34 writes *“Feature importance says something about the impact of a (changing) feature on the outcome of a prediction model in machine learning.”*

An *ablation-based* perspective was indicated by mentioning model performance when leaving out a feature. 9 of our 33 participants used words that indicated this perspective. For example, participant 4 mentions that *“Without the feature, the training metric is worse.”* and participant 14 writes that feature importance is a *“[...] indicate how much impact input variables have at the prediction of the target variable. I.e. if we remove the input, how will the prediction error increase?”*

T4-9. Purpose. We bundled the six themes that cover the purpose of feature importance: *Understanding, Feature selection, Debugging, Trust and fairness, Decision-making* and *Improve model performance*. These themes are in line with motivations for XAI described in the literature, such as social acceptance, managing social interactions, detecting faulty model behavior (debugging, auditing), and acquiring new knowledge [5]. The majority of our participants was enthusiastic about feature importance, describing its perceived value and various use cases for questions 4 and 5.

The first use case is to apply feature importance for *Understanding a model or data (T4)*. This theme was mentioned by 24 participants (73%). Categories under Understanding include explainability, justifying predictions, and discover relationships. For example, participant 21 writes *“It can provide valuable insights into the working of your model.”*

The second sub-theme regarding purpose is *Feature selection (T5)*. This theme was brought up by 11 of our participants (33%). Some keywords used by our participants include removing unwanted features, removing redundant features, and the term feature selection itself. Participant 12 writes that *“If features are considered discriminatory they can be excluded from the model input. If features seem logical they may be used to better understand the dataset.”*

The third purpose is for *Debugging (T6)*, which includes specific investigations upon the model based on the Understanding from T4. For example, it includes the identification of undesirable or unexpected behavior, the validation of the model, and understanding or preventing failure. This theme was introduced by 14 of our participants (42%). Participant 11 described: *“I have used feature importance (gradcam) in an image classification task, to see if the model was activating on the “right” parts of the image (so debugging).”*

The fourth sub-theme for purpose is to use feature importance for *Trust and fairness (T7)*. This is a broader theme, and includes using feature importance to increase trust in the model and to identify unfairness. 9 participants introduced purposes in this category (27%). Participant 16 predicts: “[...] feature importance will become part of the normal process to make sure the models are fair and does not have biases affecting customers”

The fifth use case identified by our participants is to support *Decision-making (T8)*. This theme was mentioned by 8 of our participants (24%). Participant 22 writes: “As an end user, it can help to decide how much trust to place in the outcome of a machine learning model and how to act based on that outcome.” and adds how it can speed up the decision-making process: “[...] feature importance can be used to guide an investigation towards the factors that the model found important. This may save time, because the investigation can be targeted from the start.”

The final identified purpose of feature importance is to *Improve model performance (T9)*. This theme includes performance in the sense of accuracy, energy efficiency and speed. Improving performance was brought up by 4 of our participants (12%). For example, participant 23 writes: “It can also help identify which features can be removed without affecting performance making the model more time/energy efficient.”

T10. Downsides. The final theme contains the downsides of feature importance as identified by our participants. Our participants were positive about feature importance techniques, but seven participants also reported some doubts. In general, the downsides regard the incompleteness of (the definition of) feature importance: it is unclear what score is ‘good’, it does not explain the why of a score, and it requires domain knowledge to interpret. Furthermore, one participant noted that they did not understand how to interpret local feature importance scores. Finally, one participant mentioned that feature importance scores can be misleading. Participant 4 was rather critical: “I have yet to hear about a magic number that properly captures what the actual effect of a feature is. In my mind, reality is much more nuanced. Assigning a single number to ‘importance’ is bound to lead to a misinterpretation somewhere down the line.”

Summary

To answer **RQ1**, our qualitative analysis showed that the way data scientists define feature importance varies widely. Without context,

the majority of our participants (20 of 33) see feature importance as a global technique (**T1**). Participants also mentioned various explanandums (**T2**): they argued feature importance explains the quality, informativeness and predictions of the model (each have different semantics). Furthermore, in **T3** we found that, for those participants that indicated assumptions of an underlying mechanisms, these perspectives were held equally, with 8 participants for gradient-based, and 9 for ablation-based. Finally, our participants indicated aspects that made feature importance valuable (**T4-9**), as well as downsides of the techniques (**T10**), especially that it is incomplete. A main problem is that there are no guidelines on what feature importance scores are ‘good’ or ‘bad’.

7 QUANTITATIVE RESULTS

We now turn to the quantitative results. For each of the properties identified in Section 4, participants were presented with a set of statements and asked to indicate their level of agreement. In order to elicit expectations in absence of specific implementation details, the presented examples were relatively abstract. As a result, several participants reported it was challenging to indicate their agreement. Some remarked the interpretation depends on which model is explained, or that the questions were not sufficiently specific.

P1. Actionability. The six statements related to actionability evoked the most neutral answers out of all questions (Figure 2). Specifically, 65% of all participants answered question 3 and 4 with neutral, and for the last two questions that goes up to 74%.

The first and second questions are opposites. Although the results of the first question are not significantly different from neutral, for the second question we see a slight tendency towards agreement ($p = 0.019^*$). This indicates that some experts expect feature importance to be actionable: it should indicate how instance perturbations will affect the model’s score. This corresponds to a gradient-based rather than ablation-based interpretation of feature importance.

For the other four statements, there was no statistically significant (dis)agreement. Participant 19 clarified: “3/4/5/6 are obvious, you just do not know.” and participant 34 said “5/6: same reasoning as 3/4, we don’t know if the effect is positive or negative.” This is interesting, as existing gradient-based techniques such as LIME actually do indicate the direction of change based on whether the feature importance value is positive or negative.

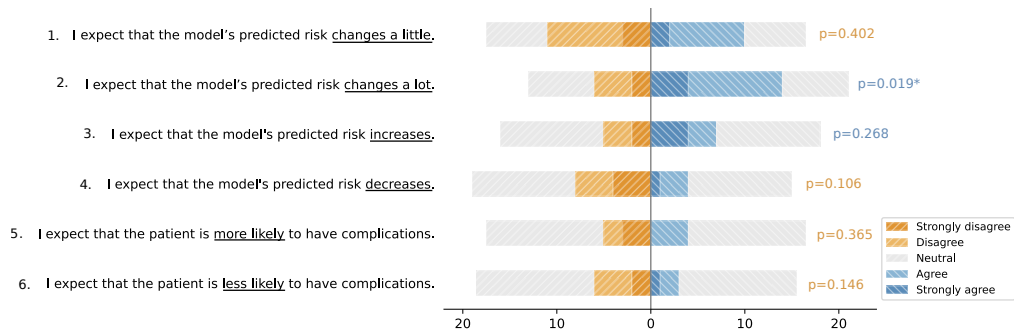


Figure 2: Is feature importance actionable?

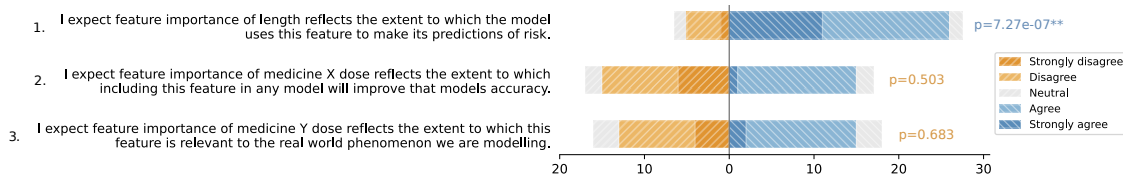


Figure 3: Is feature importance not causal?

P2. Causality. This set of questions (Figure 3) includes statements on whether feature importance explains a specific model, the informativeness of a feature in general, or the real world phenomenon. Unintentionally, this corresponds quite well with the explanandum theme (T3) from the qualitative analysis.

Overall, most participants (76%) expect that feature importance reflects the usage of a model to make predictions, as expected this corresponds well with currently available techniques. Alarmingly, there were also quite a few participants (44%) that expected feature importance to reflect how relevant the feature is to the real world phenomenon (a causal statement) which is unexpected. This does not correspond with current feature importance techniques, which all explain what caused the model to make certain predictions (correlation), as opposed to what caused the phenomenon in the real world (causation). In fact, as machine learning models are only able to identify correlation, not causation, it would require a totally different approach to computing feature importance (e.g., causal inference). As participant 34 puts it: *“Predicting how many ice creams a supermarket will sell based on how many hours I slept doesn’t make sense ... even though feature importance might be high”*. Two of the participants expecting feature importance to explain causation, did mention in the optional comment field they did not quite understand the question.

P8. Sampling Distribution. The statements in Figure 3 also relates to the sampling distribution property (P8). If feature importance strictly explains predictions (question 1), an interventional distribution is sufficient to match data scientists’ expectations. If feature importance explains the informativeness to a model or the real world phenomenon (question 2 and 3 respectively), a conditional distribution seems more appropriate. Our results only show significant evidence for the first interpretation, meaning interventional sampling for instance perturbations has most support.

P3. Stability. Most participants (68%) expected that feature importance is stable (Figure 4.1): slight changes in parameters will not impact the explanation significantly. Participant 26 justifies: *“I don’t want to spend a lot of time tuning the explanation technique.”*

Some participants remarked that, even though ideal, it may not be possible to satisfy this property for all models (especially in the case of correlated features).

P4. Robustness. Even more participants expected feature importance to be similar (Figure 4.2) for two similar data points (82%; the most agreement out of all statements). This is surprising, since this property constrains feature importance in its ability to closely match the reference model. In particular, if the reference model’s score does change rapidly, this property prevents the feature importance technique to convey the true behavior of the model. Participant 27 notes there is a difference between *“what I would expect if I were a layman [and] knowing that there can be abrupt boundaries in the input/output space.”* Overall, it seems experts favor the robustness of the explanation in spite of possible problems regarding faithfulness with respect to the reference model.

P5. Selectivity. The results show a slight tendency towards selective explanations: 56% of participants expected feature importance to *not* include all features (Figure 4.3), versus 35% favoring all features to be included. However, expectations varied a lot, and only three participants answered neutral. As preferences for selectivity seem subjective, we should strive for a more flexible approach.

P6. Additivity. Next, we analyzed participant expectations of the potential side-effects of additivity (Figure 5). As described in Section 4, additivity can potentially lead to misleading interpretations: as the feature importance values need to add up to the prediction, we need to make a decision on how to divide the importance over strongly correlated features. From the first two statements, we see participants expect the importance of a feature to include all of its interactions with other correlated features (74% and 65% respectively). Participant 11 remarked: *“I want this to be the case for a perfect explainer, [but it is] not necessarily how I think [existing methods] work.”*

In the last two statements, we gave a concrete example of two correlated features (length and weight), and two uncorrelated features (medicine X and Y dose). Here participants were a bit more

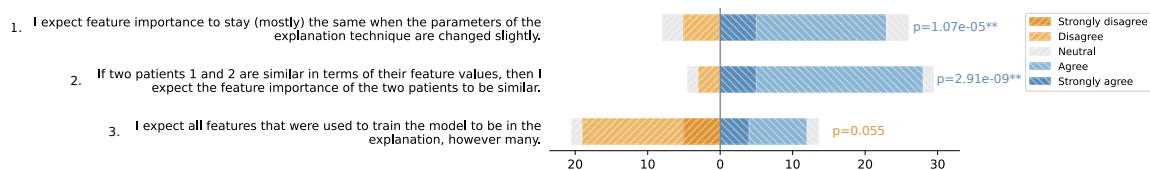


Figure 4: Is feature importance 1) stable, 2) robust and 3) selective?

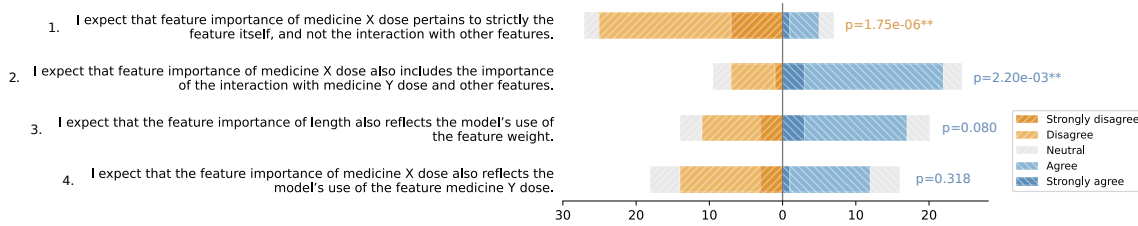


Figure 5: How is importance distributed across correlated features (consequence of additivity)?

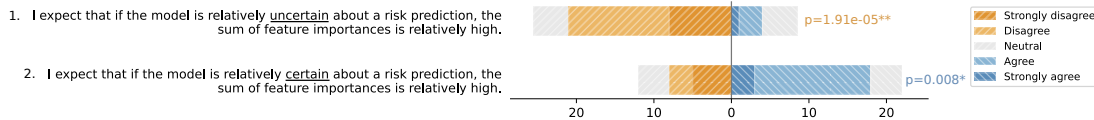


Figure 6: Is feature importance proportional?

divided: only a small majority expected feature importance to cover all correlated features (50% agreed, 32% disagreed, $p = 0.080$).

It is noteworthy that, although the importance of a feature never includes the importance of a completely unrelated feature, 35% of participants nonetheless agreed with question 4.

P7. Proportionality. Most participants clearly expected (Figure 6) feature importance to be proportional, which means that the sum of feature importance values is proportional to the model's predicted score. This matches an ablation-based feature importance perspective, and is not compatible with a gradient-based perspective.

This expectation is problematic, as the proportionality property is directly at odds with actionability: it is impossible to satisfy both properties at once (apparent in Figure 1). Yet, quite a few participants (24%) agreed with both the second statement *and* with the actionability statements in the previous set of questions. This reveals an incompatibility in data scientists' expectations.

Summary

To answer RQ2, our quantitative results indicate that data scientists mainly expected feature importance to be robust (P4, 82%), not causal (P2, 76%), additive (P6, 74%) and stable (P3, 68%). The expectation of participants varied quite a bit for the properties: selective (P5, 56%), proportional (P7, 53%) and actionable (P1, 41%). This highlights the importance of understanding the properties of different feature importance techniques, since there is no obvious choice of property that aligns with data scientists' expectations.

None of the expected properties seem to fully match existing gradient-based or ablation-based feature importance definitions. While current techniques are not causal, and mostly additive, they are generally not stable. Furthermore, robustness differs per technique: the model-agnostic techniques participants were most familiar with do not have this property [2]. Finally, gradient-based techniques are inherently actionable, but not proportional, while ablation-based techniques are not actionable, but are proportional to the model output.

8 DISCUSSION

Related work has identified and critiqued various properties of feature importance techniques. However, to the best of our knowledge, we are the first to verify the relevance and alignment of these properties with data science practitioners: one of the stakeholders that explanation techniques are ultimately meant to support.

8.1 Properties Expected by Data Scientists

In our qualitative study, our participants mention purposes that are in line with the ones described in literature: social acceptance, managing social interactions, detecting faulty model behavior (debugging, auditing), and acquiring new knowledge [5]. Interestingly, feature selection was mentioned often (33.3%), even though this goal is not often explicitly mentioned in recent literature on XAI.

The results of our quantitative study indicate that several properties of local feature importance techniques were largely expected by our participants: robustness, (non-)causality, additivity, and stability. Although several of these properties align with existing techniques, others are currently not supported.

8.1.1 Robustness & Stability. The majority of participants expected feature importance to be robust (P4, 82%). Similarly, we saw strong evidence experts expect techniques to be stable (P3, 68%). Importantly, perturbation-based feature importance approaches exhibit neither robustness nor stability if the number of samples is too small [2, 12], revealing a potential mismatch of expectations and practice. An interesting direction of future research would be to improve sampling techniques to satisfy these two properties. In particular, future work could focus on satisfying robustness without reducing the faithfulness of an explanation. Additionally, in order to manage user expectations, future work could focus on the effective communication of these inherent uncertainties.

8.1.2 Causality & Sampling Distribution. The large majority of participants expect feature importance to explain predictions (**P2**, **P8**, **T2**), supporting a non-causal interpretation of feature importance. However, some participants also expect feature importance to reflect the extent to which inclusion of the feature improves the quality of the model or reflects relevance to the real-world phenomenon varied (**P8**, **T2**). This suggests that data scientists do not expect explanations to be causal, but also do not always consider the model in isolation. The latter does not correspond with how current feature importance techniques work and could lead to incorrect conclusions about the data.

As shown in Section 4, different sampling approaches reveal fundamentally different views of what it is that feature importance explains. Our findings suggest that interventional sampling, as used by most existing techniques, may be suitable to match most data scientists' expectations - but the results are not conclusive. In particular, we believe that future work should explicitly consider the effects of the out-of-distribution problems on the (mis)interpretation of feature importance scores.

8.1.3 Additivity. Hancox-Li and Kumar [15] suggest that users may interpret feature importance to represent solely univariate effects, which would not match existing techniques. We have seen no evidence for this in our study: the large majority of participants (**P6**, 74%) expected a feature importance value to include all interactions with other features. This is consistent with existing additive techniques, such as LIME and SHAP.

8.2 Properties with Varying Expectations

For selectivity, proportionality, and actionability, expectations were much more varied, highlighting the importance of clearly communicating the underlying properties of a particular technique.

8.2.1 Selectivity. In our study, participants widely varied in their preference for selectivity (**P5**), reflecting the possible tension between the faithfulness and selectivity of an explanation. These results suggest that a more flexible approach towards selectivity is desirable over selectivity inherent to the explanation algorithm (e.g., L1 regularization in LIME). In particular, we recommend including an option to filter out features with low importance when feature importance is presented to the user.

8.2.2 Actionability & Proportionality. As explained in Sections 2 and 4, gradient-based feature importance is inherently actionable, but not proportional. In contrast, ablation-based feature importance scores are not actionable, but proportional to the model output. For actionability we observed a slight tendency towards agreement, most participants did not expect feature importance scores to be actionable (**P1**). Contrarily, most participants did expect feature importance to be proportional (**P7**). Importantly, certain participants (24%) had expectations that fundamentally contradict each other: they expected feature importance to be both actionable and proportional. In our qualitative analysis we see similar results (**T3**).

We speculate that this contradiction arises from overloading terminology of the term 'feature importance', which is insufficient for explaining what existing techniques do. To address this, we propose different terms to refer to local gradient-based and ablation-based

techniques. For gradient-based feature importance we suggest '**feature sensitivity**', as these values describe the sensitivity of the model towards changes in this features value. Next, we suggest '**feature attribution**' for ablation-based feature importance, because 'attribution' implies the additive nature of these techniques. This term has already been used by some authors, such as Lundberg and Lee [25], but not consistently. We hope that using different terms helps data scientists to recognize the differences and update their expectations of how feature importance scores should (not) be interpreted.

8.3 Limitations

In this paper, we explore the mental models of data scientists through an online survey. This method is convenient for gathering larger samples of data, but is also prone to some biases. First of all, there is the risk of selection bias, as data scientists with an above average interest in XAI (and therefore a better understanding of existing feature importance techniques) may be more likely to respond to the survey. However, we argue that targeting this audience enables us to uncover misconceptions that are held despite a good understanding of existing techniques, which tend to be more problematic due to their persistence.

Furthermore, there is a risk of response biases associated with the Likert-scale questions, such as extreme responding and primacy bias. To reduce the effects of these biases, we have carefully considered the wording of the questions between the authors, and ask each Likert-scale question in both directions (both negatively and positively framed).

This approach of checking each question both ways also makes it more likely that our questions accurately capture the participants' perspectives (interpretive validity). The two-way questioning uncovered contradictions that would be much more difficult to surface from more unstructured data such as interviews, as "[...] participants may be unaware of their own feelings or views, may recall these inaccurately, and may consciously or unconsciously distort or conceal their views." [27, p. 290].

Our sample of participants was not large enough to ensure all results fully generalize to other communities (external generalizability). However, even in a small sample, we have found significant contradictions in our data scientists' mental models of explanation techniques that are unlikely to be just outliers.

With regard to internal generalizability (generalizing within the same group to unseen examples and questions) the aim was to have generalizable questions, contextualized with an example. The introduction of this running example served to make the statements more concrete and easier to read. However, it caused some participants to (mis)interpret these statements as questions specific to the running example. For example, some participants reported they considered *length* to be relevant for medical prediction, as opposed to whether they expected any feature with similar characteristics to be important. This is inherent to the nature of our exploratory study. In future work, we may compensate for this lack of internal generalizability by conducting a more elaborate validation of the questionnaire (e.g., through pilot testing) or introducing more running examples. However, the latter would limit the number of properties that can reasonably be covered in a single study.

Finally, although these exploratory findings are a good first step towards uncovering mental models, conducting interviews could have been valuable. Although Likert-scale questions are good for quantifying opinions and uncovering contradictions, they are also a very closed-off method. In future work, it would be interesting to examine mental models of local feature importance through interviews, as those may give more specific insight into the misconceptions that the data scientists have.

8.4 Future Work

In addition to the avenues for future work discussed so far, we envision several extensions of our work. First, future work may consider data scientists' expectations of other additional properties such as faithfulness [28], contrastiveness [28], and representativeness [31].

Additionally, data scientists may have a different understanding of feature importance from domain experts or the general public. Since these groups have a lesser understanding of machine learning models and how explanations are derived, this may increase the risk of misinterpretations. For example, as opposed to the data science practitioners surveyed in this study, other groups may interpret feature importance scores univariately. Future work should consider the mental models of other stakeholders for feature importance.

This work presented the first steps in exploring the mental models of feature importance. The results of our study warrant more targeted HCI work to study individual properties in more detail.

9 CONCLUSION

In this paper, we investigated local feature importance scores that quantify the importance of the feature values to a prediction of a particular instance. While these techniques are popular and many techniques exist, they have widely varying underlying assumptions of what 'importance' means.

To address this, we surveyed related work and present an overview of several key properties of local feature importance approaches that may lead to misleading interpretations. We conducted a mixed-methods survey to explore the expectations of data scientists in industry. We found that data scientists have widely varying definitions of feature importance and its values (**RQ1**), especially regarding the themes Locality, Explanandum and Underlying mechanism (**T1-T3**). Regarding the properties of local feature importance (**RQ2**), while we found evidence that the identified properties are indeed largely expected by practitioners, data scientists also held intuitions that do not necessarily fit with existing techniques. For example, while existing techniques are not causal, and mostly additive, they are generally not stable and can lack robustness. Next, we uncovered contradicting expectations of both actionability and proportionality, which cannot be satisfied simultaneously.

We argue that this contradiction is the result of fundamental differences in how feature importance is derived (gradient and ablation-based) and should be more clearly reflected in communication about the techniques. We believe that our work can provide a fruitful starting point for future research in this direction. The uncovered expectations provide a basis for XAI researchers to further improve and explain feature importance techniques. Moreover, our exploratory study identified plenty future work for HCI researchers in unraveling users' mental models of XAI.

ACKNOWLEDGMENTS

We thank all of our participants for their time and valuable feedback. This work is part of the research programme Commit2Data, specifically the RATE Analytics project with project number 628.003.001, which is financed by the Dutch Research Council (NWO).

REFERENCES

- [1] André Altmann, Laura Tološi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics* 26, 10 (2010), 1340–1347.
- [2] David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *ICML Workshop on Human Interpretability in Machine Learning* (2018), 66–71.
- [3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11, Jun (2010), 1803–1831.
- [4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.
- [5] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.
- [6] Ashley Castleberry and Amanda Nolen. 2018. Thematic analysis of qualitative research data: is it as easy as it sounds? *Currents in Pharmacy Teaching and Learning* 10, 6 (2018), 807–815.
- [7] Hugh Chen, Joseph D Janizek, Scott Lundberg, and Su-In Lee. 2020. True to the model or true to the data? *ICML Workshop on Human Interpretability in Machine Learning* (2020), 123–129.
- [8] Furui Cheng, Dongyu Liu, Fan Du, Yanna Lin, Alexandra Zytek, Haomin Li, Huamin Qu, and Kalyan Veeramachaneni. 2021. VBridge: Connecting the dots between features and data to explain healthcare models. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 378–388.
- [9] Dennis Collaris and Jarke J van Wijk. 2020. ExplainExplore: visual exploration of machine learning explanations. In *2020 IEEE Pacific Visualization Symposium*. IEEE, 26–35.
- [10] Dennis Collaris, Leo M Vink, and Jarke J van Wijk. 2018. Instance-level explanations for fraud detection: A case study. *ICML Workshop on Human Interpretability in Machine Learning* (2018), 28–33.
- [11] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [12] Damien Garreau and Ulrike Luxburg. 2020. Explaining the explainer: A first theoretical analysis of LIME. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1287–1296.
- [13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *Comput. Surveys* 51, 5 (2018), 1–42.
- [14] David Gunning and David Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine* 40, 2 (Jun. 2019), 44–58. <https://doi.org/10.1609/aimag.v40i2.2850>
- [15] Leif Hancox-Li and I Elizabeth Kumar. 2021. Epistemic values in feature importance methods: Lessons from feminist epistemology. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 817–826.
- [16] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [17] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26.
- [18] Giles Hooker, Lucas Mentch, and Siyu Zhou. 2021. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing* 31, 6 (2021), 1–16.
- [19] Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. 2021. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 805–815.
- [20] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [21] Igor Kononenko et al. 2010. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research* 11, Jan (2010), 1–18.
- [22] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages*

- and *Human Centric Computing*. IEEE, 3–10.
- [23] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*. PMLR, 5491–5500.
- [24] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
- [25] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 4768–4777.
- [26] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2, 10 (2018), 749–760.
- [27] Joseph A. Maxwell. 1992. Understanding and Validity in Qualitative Research. *Harvard Educational Review* 62, 3 (1992), 279–300.
- [28] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
- [29] Donald A Norman. 1983. Some observations on mental models. In *Mental Models*. Psychology Press.
- [30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.
- [31] Marko Robnik-Šikonja and Marko Bohanec. 2018. Perturbation-based explanations of prediction models. In *Human and Machine Learning*. Springer, 159–175.
- [32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
- [33] Erik Štrumbelj, Igor Kononenko, and Marko Robnik-Šikonja. 2009. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering* 68, 10 (2009), 886–904.
- [34] Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A human-grounded evaluation of SHAP for alert processing. *KDD Workshop on Explainable AI* (2019).
- [35] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.
- [36] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. Springer, 818–833.
- [37] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. “Why should you trust my explanation?” Understanding uncertainty in LIME explanations. *ICML AI for Social Good Workshop* (2019).