Info Vis

Article

Information Visualization
1–19

DOI: 10.1177/14738716261434842
journals.sagepub.com/home/ivi


Evaluating the utility of feature importance visualizations in SHAP

Dennis Collaris[1], Yu Liang[2], Martijn C. Willemsen[2], Angelos Chatzimparmpas[1] and Jarke J. van Wijk[2]

Abstract

Feature importance is a technique that helps users understand machine learning models by showing how much each feature contributed to the model's predictions. For example, it can be used for housing price prediction to explain why certain features lead to higher or lower prices. Different visualizations are used to convey feature importance to users: standard bar charts, but also advanced waterfall plots and force plots as provided by SHAP. These advanced visual representations convey more information (e.g., about the additivity property of the technique). However, this may come at the expense of the figure's simplicity. Both the trade-off between these properties and the added benefit of these advanced visualizations have yet to be formally studied. In this paper, we evaluate the effectiveness of three common SHAP visualization types for users to understand how machine learning-based prediction works in a housing price prediction scenario. Each participant answered a set of quiz questions aimed to measure their basic understanding of the feature importance (the absolute impact of features), the negative or positive impact of features, and the additivity property of feature importance. By testing whether participants understood these concepts, we assert whether the advanced visual metaphors are effective in conveying additional information beyond the standard bar chart visualization. Moreover, we study whether the effectiveness is moderated by personal characteristics, such as an individual's visual familiarity and cognitive skills. Our results from 2 user experiments comprising 546 participants in total show that, despite testing specifically for the properties that waterfall plots emphasize, bar and waterfall plots perform equally well. Force plots seemed to perform worse, and these results were independent of the skills and experience of the participants. Therefore, our findings tentatively suggest that bar charts may be a preferable choice for communicating feature importance due to their simplicity and comparable effectiveness.

Keywords

explainable AI, feature importance, visualization, Shapley values, comparative user experiment

Introduction

As Machine Learning (ML) models become more prevalent in decision-making processes, it is crucial to ensure that they are trustworthy and accountable. This requires the development of techniques that enable us to understand how these models make their predictions. A popular approach to explain individual predictions by ML models is feature importance techniques, which assign a scalar value to each feature to indicate its importance to the prediction for an individual data point. One widely adopted approach is SHAP[2] (SHapley Additive exPlanations), which quantifies each feature's contribution to an individual prediction. As an example, such techniques can be used for housing price prediction to explain why certain

[1]Utrecht University, Netherlands
[2]Eindhoven University of Technology, Netherlands

Corresponding author:
Dennis Collaris, Department of Information and Computing Sciences, Utrecht University, Utrecht 3584 CC, Netherlands.
Email: d.a.c.collaris@uu.nl

Figure 1. Three common visualizations of feature importance for a price prediction of a house in the Boston housing dataset.[1]

features lead to higher or lower prices. By understanding the importance of features visually, non-expert users can reason about which property characteristics (e.g. location, size, or energy efficiency) drive a predicted price and use this information to support practical decisions, such as comparing properties, assessing value, or deciding how much to bid.[3] Therefore, feature importance visualizations should be designed to be understandable and to support concrete actions,[4] even for users who interpret them primarily as intuitive summaries of how input features influence model outcomes. In practice, these visualizations are often used by individuals without formal training in ML or explainable AI (XAI), who rely on them for understanding and communication rather than for engaging with their underlying mathematical foundations. Ultimately, feature importance techniques aim to enable more trustworthy and accountable use of ML models.

At first glance, these techniques seem like an intuitive and simple solution to explain ML models. However, there are many pitfalls that practitioners without an AI background can encounter when using these techniques. For example, the notion of "importance" relies heavily on the underlying mechanism used to infer that score, and varies a lot between different techniques. Different techniques may exhibit different properties, which may not correspond with the expectations of the users working with these scores.[5]

The visualization of feature importance is an important aspect of helping users understand models. To this end, different visualizations are used to convey feature importance to users. In particular, SHAP explanations are commonly visualized using bar charts, waterfall plots, and force plots. The most straightforward approach is to visualize the weights of each feature using a **bar chart** (Figure 1↖). However, advanced visual metaphors have also been introduced to convey additional information, such as implicitly illustrating the additivity property of the technique. For example, **waterfall plots** (Figure 1↗) imply a specific ordering and emphasize additivity by consecutively aligning the bars and connecting them with lines. Next, another visual approach is the **force plot** (Figure 1↓), bundling the positive and negative contributions to either side of the prediction, and implying a "pushing force" from the base rate threshold toward the final prediction value.

With the huge popularity of feature importance techniques for explainability, the use of these visualizations has quickly become ubiquitous. However, the effectiveness of these visualizations to help users understand feature importance has not formally been

studied. While the advanced visualizations introduce more information, it is unclear whether the cost of adding more complexity to the visualizations is outweighed by the benefit of the additional information toward understanding the model. Meanwhile, it is critical that the validity of visualization techniques is assessed,[6,7] yet no such assessment exists.

To assert the validity of these feature importance visualizations, we evaluate the effectiveness of these visualizations using a realistic housing price prediction scenario. Each participant is first provided with one of the two visualizations randomly assigned to them. Next, they answer a set of multiple-choice questions aimed at measuring their understanding of (1) the absolute impact of features; (2) the positive or negative impact of features; and (3) the additivity property of feature importance. After answering these questions, they were prompted to reflect on their experience. This process is then repeated for a second visualization. By testing whether participants understood these concepts, we can assert whether the advanced visual metaphors are effective in conveying additional information beyond the standard bar chart visualization. Furthermore, we study whether the effectiveness is moderated by personal characteristics, such as an individual's visual familiarity and cognitive skills. Finally, we conclude by providing recommendations on which visualizations to use for feature importance visualization.

Background and related work

Explainable AI visualization

In the visualization community, many Visual Analytics (VA) systems have been proposed to support the understanding of ML models.[8–11] We instead focus on simpler building blocks that such systems are typically comprised of (for an extensive review of XAI methods, see Molnar's book[12]). These are easier to evaluate and test in isolation, and our findings can still be mapped back to the larger VA systems (e.g.,[13–15]).

One such building block – and the focus of this study – is feature importance visualization. Feature importance in general is the quantitative assignment of importance or influence to the features used by an ML model. A recent survey paper by Chatzimparmpas et al.[8] identified this area as a research opportunity for the visualization community, as only a limited number of works, such as those by Angelini et al.[16] and Kerrigan et al.,[17] have explored steerable and enhanced *global* feature importance visualizations, including Partial Dependence Plots (PDPs), to support the analysis of complex models and high-dimensional feature spaces. However, in this study, we focus on *local* feature importance, which is concerned with the influence of features toward a single prediction of a model,[18] and only explanations for tabular data.

Feature importance visualization

Local feature importance is by far the most popular explanation technique due to its seemingly simple interpretation and scalability.[19] However, this interpretation of "importance" hinges on the underlying mechanism used to infer that importance, and can vary a lot between different techniques.[5] This may result in practitioners having unrealistic expectations of what these techniques do. Visualization is an important aspect of helping users understand models with feature importance. It helps convey the right message and can put emphasis on certain properties of the feature importance scores through its visual encoding. Visualization design choices are known to strongly affect how viewers perceive magnitude, order, and causality in charts, as demonstrated by decades of research on graphical perception and crowd-sourced evaluation of visual encodings (e.g.,[20–22]).

Local Interpretable Model agnostic Explanations (LIME) popularized this feature importance approach in 2016.[23] To generate LIME explanations, we first sample new observations from around the instance being explained, and obtain their predictions using the original model. We then weigh these samples based on their proximity to the target observation and use them to construct a linear surrogate model that approximates the original model. The coefficients of that local surrogate can be used as feature importance scores and can be interpreted as the rate of change in predicted output when changing the feature value (e.g. an approximate derivative of the model's outcome function, or the feature *sensitivity*).

LIME is typically visualized using a divergent vertical bar chart (such as shown in Figure 1↖). It is important to carefully choose the positive (right) and negative (left) side, as framing based on the decision class may be counter-intuitive.[24] Bar-based encodings in particular have well-documented strengths and limitations, including sensitivity to ordering, baseline choice, and framing effects.[25,26]

SHapley Additive exPlanations (SHAP) was later introduced[2] and finds its basis in game theory. Shapley et al. values[27] find a fair distribution of the score amongst players in a cooperative game. This technique can be applied to ML predictions to figure out importance.[28] To address the exponential time complexity of the original algorithm, modern techniques[2,28] such as SHAP use a sampling approach to reduce the number

of coalitions that need to be compared. Shapley value-based feature importance can be interpreted as the proportion of the predicted outcome that can be explained by the inclusion of a feature in the model (e.g. feature *attribution*).

SHAP values can be visualized as a bar chart (cf. Figure 1↖). However, the authors adopted and popularized additional visualization approaches within XAI, most notably waterfall (Figure 1↗) and force-based plots (Figure 1↓). Waterfall plots emphasize additivity by consecutively aligning the bars and connecting them with lines, implying a specific order of application of the feature importance. One starts from the base rate (i.e. the expected average prediction from the model), and iteratively adds the impact of each feature to end up at the prediction of the model (e.g. additivity). Alternatively, force plots bundle together the positive and negative contributions to either side of the prediction, and imply a "pushing" force from the base rate threshold toward the final prediction value (e.g. additivity). Both visualizations enable users to simulate the prediction and (approximated, simplified) inner workings of the model. Waterfall plots have been used in various applications throughout the years to show how separate data values together lead to a result (e.g.,.[29,30] In the context of XAI, they have also been used to visualize the feature importance of inherently interpretable GAM models.[31] Similarly, force plots are frequently used in academic studies to present feature contributions (e.g.,[32–34]).

Feature importance visualizations are typically introduced along with feature importance technique as an auxiliary contribution,[2,23] and are rarely studied in detail. A recent study[35] shows that these visualizations used as defaults in popular XAI methods like SHAP can be confusing even for users with mathematical or statistical backgrounds and intermediate to advanced knowledge of ML and model interpretability. However, to the best of our knowledge, no formal evaluation of these visualizations has been carried out with lay users without an AI background – who may benefit the most from such explanations[36,37]– despite the ubiquity of these visualizations in XAI literature and practice.

In this work, we therefore focus on Shapley value–based explanations, as they are the most widely used (due to the method's robust guarantees) and representative of current explainability practices.[38–40] Additionally, this technique is commonly accompanied by visual encodings that repurpose established chart types (e.g. waterfall and force plots) for explaining model predictions; however, their effectiveness and potential interpretation pitfalls in XAI remain underexplored in this context.

Properties of feature importance

Prior work from Collaris et al.[5] outlined several properties that feature importance techniques can exhibit. The three properties that are most prominently represented in feature importance visualizations are "additivity," "proportionality" and "actionability."

Additivity means that the feature importance technique can be interpreted as a decomposition of the predicted score of the model over all features, resulting in one value per feature.[2] This is explicitly conveyed in waterfall and force plots, as you are supposed to read both visualizations as starting from the base rate, sequentially applying feature contributions, ending up at the final prediction.

Proportionality results from additivity and implies that the sum of feature importance values is proportional to the output of the original model. For the visualization of the technique, this means that the user can infer the final prediction solely from the feature importance values. SHAP explanations are proportional, while LIME ones are not.

Actionability means that if a feature is important, an action (i.e. a small change in feature value) will affect the model's score.[41] If a visualization conveys LIME feature importance, this is possible, whereas if SHAP is used, it is not always possible.

Our goal is to investigate the added value of explicitly conveying these properties in feature importance visualizations and whether these concepts are understood by the readers of these visualizations, regardless of their background in (X)AI.

Individual/personal differences in XAI

The extent to which feature importance visualization will be effective will not only depend on the quality and understandability of the visualization itself, but also on whether the user of the visualization can understand it and is willing to spend the cognitive effort in doing so. Recent work in intelligent tutoring systems,[42] XAI,[43] and explanations for recommender systems,[44,45] has shown that personal characteristics can play an important role in how well an explanation is understood by a specific user. For example, how effective and satisfactory a particular level of detail of an explanation is, can depend on users' need for cognition[46] or visual familiarity.[45] Similarly, the need for explanations in a recommender system was different among people with high or low need for cognition, and users' interactions with the (visual) interface also depended on the visual literacy of the user.[44] Similar to this earlier work, we also expect that the effectiveness of different visualizations in our study will depend on users' cognitive skills and need for cognition, their

visual familiarity, and how experienced they are with AI.

Research question and hypotheses

Our main goal is to validate the validity of feature importance visualizations and to provide actionable recommendations for using certain visualizations to understand machine learning models. To this end, we pose the following research questions:

RQ1. How do different visualizations influence the user's understanding of feature importance?

RQ2. How does the effectiveness of feature importance visualizations depend on personal characteristics (e.g. visual familiarity and cognitive skills)? and

RQ3. Does the users' subjective assessment of the visualizations align with their performance on the questions?

We expect that the more complex visualizations (i.e. force and waterfall plots) are more effective in conveying feature importance to our participants; hence, we expect to confirm an added benefit for these visualizations. Next, we expect that more expert users (characterized by higher visual literacy, higher cognitive skills, and/or experience with AI) will benefit from more complex visualizations, whereas novice users may prefer simpler visualizations.

Methodology

Use case

As context for our explanations, we used a housing price prediction use case. In particular, we first built an ML model to predict the price of a house in Boston. The algorithm was trained on a dataset[1] collected by the U.S Census Service concerning housing (from the '70s) in the area of Boston, Massachusetts, based on a set of 12 house features: crime rate, % residential zone, % industrial zone, Charles River, NOX concentration, number of rooms, % built before 1940, remoteness, connectedness, tax rate, pupil-teacher ratio, and % working class. We used a gradient boosted trees model (*lightgbm*) as the complex model we would like to investigate, and selected two random houses, A and B, from the dataset to generate explanations for.

Study design

The study was conducted using a mixed factorial design. The three different visualizations (bar chart, force plot, and waterfall plot) were compared both between and within subjects. Across conditions, all pairwise combinations of the three visualizations were tested. Participants were assigned two randomly chosen visualizations, and for each, they were asked to answer a set of questions on feature importance, considering the price prediction for a single house (from two different scenarios: house A and house B). This pairwise comparison design reduces cognitive burden compared to multi-way comparisons and enables clearer relative judgments between visualization alternatives.[47] Moreover, limiting the number of conditions per participant helps reduce fatigue and maintain attention throughout the study.[48]

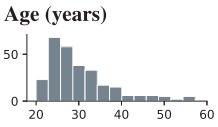
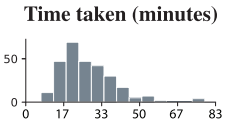
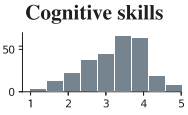
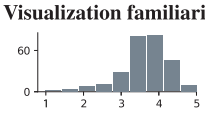
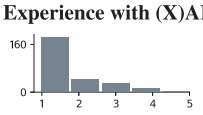
The order of the visualizations shown to the users was counterbalanced to ensure that we could not only conduct a between-subjects comparison on the first presented visualization, but also a within-subject comparison between the first and second presented (after controlling for carry-over effects). Additionally, we accounted for differences between the two houses being explained (house A and house B) by swapping the order of the explained houses for half of the participants as well.

The conditions randomly assigned to participants (both in original order and counterbalanced order) were:

C1. the **Bar chart** and **Force plot**;
C2. the **Bar chart** and **Waterfall plot**;
C3. and the **Force plot** and **Waterfall plot**.

As the differences between the different visualizations may be small, and because we also want to check for individual differences, we selected a relatively large pool of participants in order to have sufficient statistical power to draw conclusions from. Based on power analysis ($\beta = 0.9$, $\alpha = 0.05$), we need around 288 participants to be able to show medium effect sizes in comparing our three conditions (Power analysis based on an independent t-test, effect size $d = 0.5$, alpha corrected for three comparisons between the conditions). Aiming for sufficient statistical power inherently limits the number of conditions we can test for, as each new condition would increase the required number of participants by about 100 participants. We therefore chose to test for the most ubiquitous visualizations for local explanations first. Later on, we added more conditions in a follow-up study to investigate the design decisions for these visualizations (see "Follow-up study" section). For completeness, in all statistical models reported in the "Results" and "Follow-up study" sections, the bar chart condition was used as the reference (baseline) level.

Table 1. Summary of the main study participant demographics.

Type	Answers (count)
Sex	Female **(165)**, Male **(119)**, Prefer not to say **(0)**
Nationality (38 total)	Europe **(189)**, Africa **(70)**, North America **(12)**, South America **(5)**, Asia **(7)**
Age (years)	
Time taken (minutes)	
Cognitive skills	
Visualization familiarity	
Experience with (X)AI	

Study procedure

At the beginning of the study, participants were presented with an informed consent form that explained the purpose and procedure of the study. After providing informed consent, they were presented with a survey on their cognitive skills, visual familiarity, and knowledge of AI. Participants were then introduced to the housing prediction scenario. In particular, they were told an ML model was trained to predict the housing price in Boston based on the features. Also, a description of the 12 features was provided. Next, participants moved to the first scenario, in which they were asked to consider the price prediction for a single house A. They were shown one type of feature importance visualization (based on their randomly assigned condition) and were asked to use the visualization to answer a set of questions designed to measure their understanding of the feature importance. After answering, participants were asked to fill in a survey on their subjective experience with the visualization.

Next, the participants were presented with the second scenario, in which they considered another house, B, and were shown the second feature importance visualization from their assigned condition. They again filled in some questions measuring their objective understanding, and afterward were asked to reflect on their subjective experience.

Participants

We recruited participants through the Prolific platform, and restricted candidates to having a sufficient mastery of the English Language, no visual impairment, and at least a bachelor's degree. As for the reimbursement, participants were paid £3.00 for 25 min of participation. The median completion time for participants was 24:57. The study was approved by the Ethical Review Board of our institutions and complied with GDPR. In total, we ended up with 284 respondents, but had to exclude a few of these from our analysis due to their failure to respond correctly to the attention check questions. This resulted in a total of 267 participants. More details about the participant sample are shown in Table 1; the full, raw demographic data collected from participants are available in the Supplemental material.

Measurements

To evaluate how individual differences play a role in the effectiveness of visualization techniques (RQ2), we first measure participants' personal characteristics using established sets of questions introduced in prior work on visual familiarity,[45,49] cognitive skills,[44,46,50] and experience with AI[43] (cf. Table 1, last row).

To evaluate the effectiveness of different visualizations for feature importance, users were presented with a type of visualization (bar charts, force plot, or waterfall plot) designed for showing the feature importance of the prediction model for housing prices. Using the visualization, they were asked to answer nine feature importance questions.

The mentioned visualizations can be seen as a progression from simple to more complex, where each step introduces additional information in the form of (1) the positive and negative aspects of the correlation to the prediction, and (2) explicitly representing the additivity property (i.e. the combined effect results in the prediction). This is illustrated in Figure 2. To evaluate whether this additional information in the visualizations is understood by our participants (RQ1), we crafted performance questions in three categories:

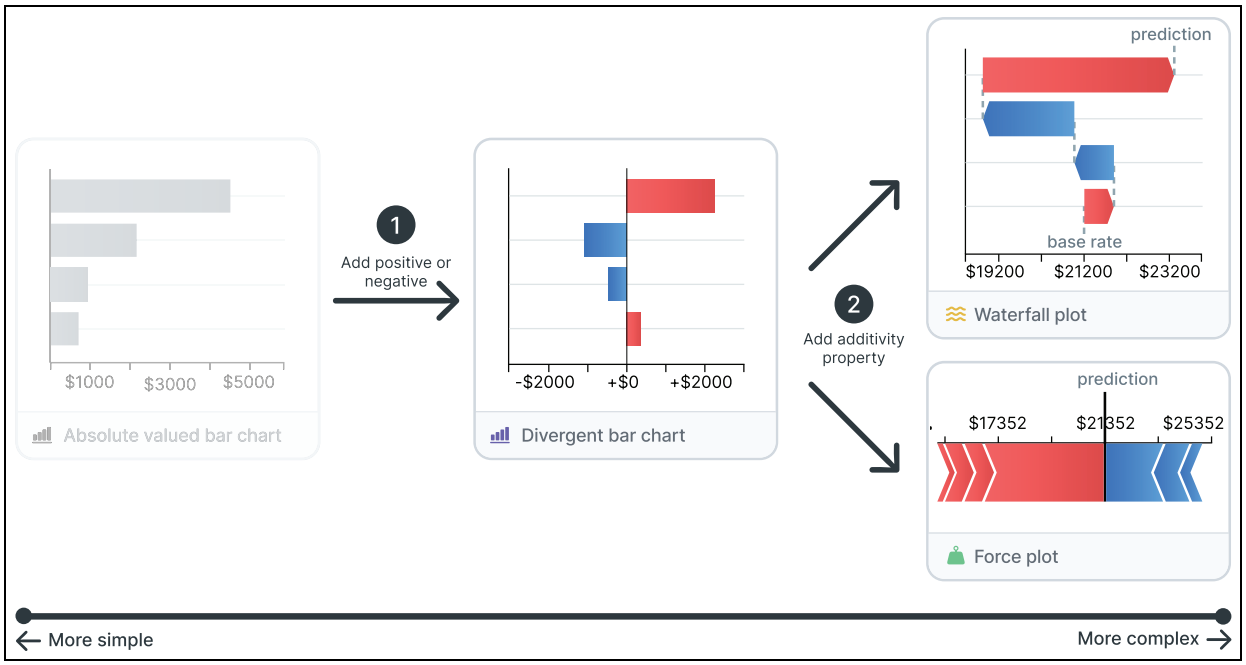


Figure 2. From simple to more complex feature importance visualizations by adding (1) the positive and negative aspects of the correlation to the prediction and (2) explicitly representing the additivity property (i.e. the combined effect results in the prediction).

- Two questions designed for measuring users' basic understanding of the feature importance (e.g. "*Which of the following statements is correct? The feature X is among the top three most important contributors to the predicted house price.*");
- Four questions for measuring users' understanding on positive and negative impact of features (e.g. "*Which of the following statements is correct? The feature X has a positive/negative impact on the predicted house price.*"); and
- Three questions for users' understanding on additivity (e.g. "*Features X and Y have a combined effect on the house price of…*").

The order of these performance questions was randomized for each participant to further avoid carry-over/learning effects. The objective effectiveness of the visualizations was measured with task correctness and response time on these questions.

After the performance questions, we also measured participants' subjective experience (RQ3), again using established sets of questions introduced in prior research on perceived informativeness[44,51–55] and effectiveness.[51,52,56,57] These were answered on a five-point Likert scale. The performance and experience questions were then repeated for the second visualization from the condition assigned to the participant.

See the Appendices of this paper for the full list of questions asked in our study. The code and all materials required to replicate our study are provided in the Supplemental material.

Results

Objective-measure questions (RQ1)

To study how the different visualizations support user understanding, we first compare the participants' performance on the three types of questions (basic understanding, positive and negative impact of features, and understanding of additivity), which were distributed across nine quiz questions for each visualization. Figure 3 shows that, on average, correctness was highest for the basic understanding questions (Q1 and Q2), and lower for the other questions, with surprisingly low scores for Q6 and Q8 (discussed in detail in the "Discussion" section). We note that across the questions, we see the lowest scores for the force plot, whereas bar and waterfall plots seem to score quite similarly.

To test for differences in understanding between visualizations, we applied multilevel models on the correctness score per question, testing the effects of the visualization, type of question, type of house, and order of visualizations on user correctness (since user

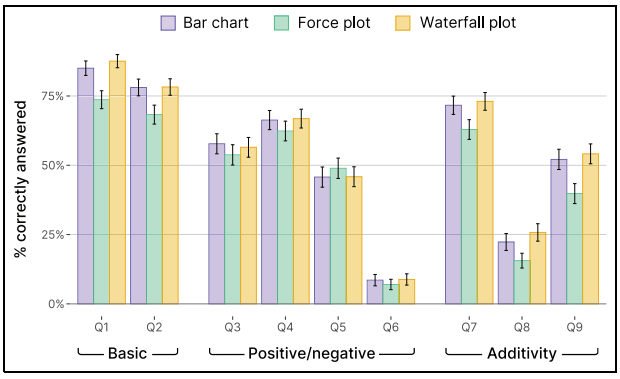


Figure 3. Mean scores for the nine questions across the three visualizations.
Error bars are one standard error of the mean.

correctness is binary for each question, logistic regression is used for the multi-level model. The model is fit with a binomial generalized linear mixed model by maximum likelihood with Laplace Approximation). As we have nine questions per visualization and participants assessed two visualizations, we have a total of 18 observations. We considered the questions per question type (basic, positive/negative, additivity) rather than per individual question to focus on the added information between the different visualizations.

Inspecting initial regression models, we find significant effects of the different visualization types, with some differences across different types of questions and some effects of order, and no differences between the house predictions. We present the results of the final model that includes the order, as (small) learning effects show interesting differences across the visualizations. In this final model (see Table 2), the baseline is the bar chart visualization and the basic question type. The effect of order is centered, such that the main effects of the conditions are averaged across the two orders (the second visualization is positive).

Force plot performs worse than the other two visualizations. In Figure 4, we show the correctness per question type and across scenario order for each of the three visualizations. Inspecting the regression and the figure, we observe two strong effects. First, positive/negative and additivity question types score lower in correctness than the basic questions ($\beta_{posneg} = -2.18, p < .001$ and $\beta_{additivity} = -1.98, p < .001$), suggesting these questions were more difficult for

Table 2. Multilevel logistic regression on correctness per visualization, question type, and order.

Visualization	Question type	Order	Correctness
(Intercept)	-	-	1.94 (0.17)***
force	-	-	−0.74 (0.21)***
waterfall	-	-	0.09 (0.23)
-	pos/neg	-	−2.18 (0.18)***
-	additivity	-	−1.98 (0.19)***
-	-	second	0.89 (0.34)**
force	pos/neg	-	0.74 (0.24)**
waterfall	pos/neg	-	0.03 (0.25)
force	additivity	-	0.37 (0.25)
waterfall	additivity	-	0.21 (0.26)
force	-	second	−0.57 (0.45)
waterfall	-	second	0.91 (0.48)
-	pos/neg	second	−0.96 (0.36)**
-	additivity	second	−1.16 (0.37)**
force	pos/neg	second	0.91 (0.48)
waterfall	pos/neg	second	1.48 (0.51)**
force	additivity	second	1.37 (0.50)**
waterfall	additivity	second	1.37 (0.52)**
AIC			5811.6
BIC			5934.6
Log Likelihood			−2886.8
Num. obs.			4794
Num. groups: userid			267
Var: userid (Intercept)			0.87
Marginal/Conditional R^2			0.138/0.320

The order is centered and is positive for the second visualization shown.
***$p<0.001$; **$p<0.01$; *$p<0.05$.

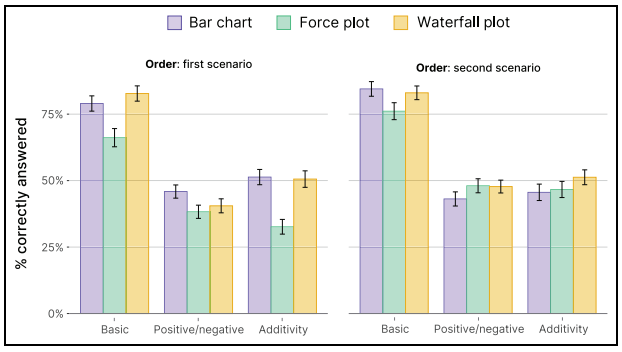


Figure 4. Correctness for each visualization across the three types of questions and order.
The force plot performs worse than the other two visualizations (especially for Basic and Additive questions), especially when the force plot was shown first.

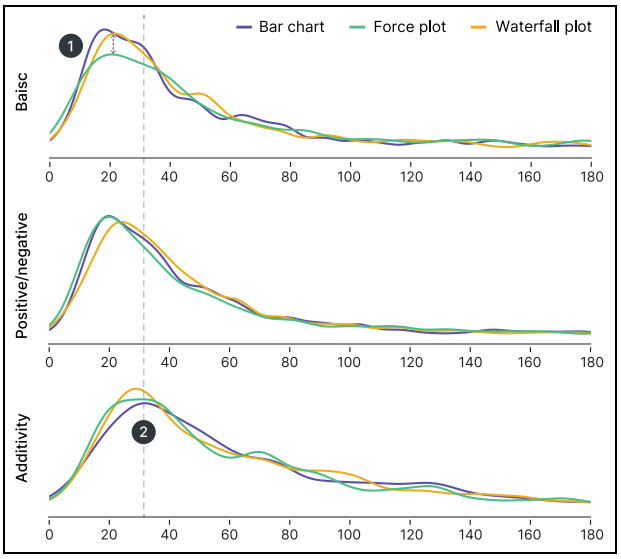


Figure 5. Response time (in seconds) for each visualization across question types.
We see (1) basic questions took longer with force plots, and (2) additivity questions took slightly longer than the other question types.

participants. Between the three visualizations, especially for the first scenario, we find that the force plot scores lower than the bar chart. This holds especially for both the basic and additive questions, and to a lesser extent for the pos/neg question types. This is supported by the regression model that shows a main effect of force plot ($\beta_{force} = -0.74, p < .001$), reflecting that for the basic questions (baseline), the force plot does worse. However, the positive effect ($\beta_{force:posneg} = 0.74, p < .01$) counteracts this effect (This effect is additive to the main effect of β_{force}.) for the pos/neg questions, whereas the effect ($\beta_{force:additivity} = 0.37, n.s.$) is small and not siginficant, and thus not counteract the negative effect of force for additivity questions. In Figure 4, we see that these differences do not appear for the second scenario: in general performance goes up for the second scenario ($\beta_{second} = 0.89, p < .01$) and further interactions of second with force and pos/neg ($\beta_{force:pos/neg:second} = 0.91, p.06$) and ($\beta_{force:pos/neg:second} = 0.91, p < .01$) show the lower perfomance in the first scenario of force plots, disappears in the second scenario, suggesting some learning from the first to the second scenario. Note that participants who get a force plot in the second scenario have seen a bar chart or a waterfall plot in the first scenario, and perhaps encountering these other visualizations helps them better understand the force plot.

Hardly any difference between bar chart and waterfall plot. For the first scenario, our model does not show any significant difference between the waterfall plot and bar chart, as all effects that include the waterfall plot (which is compared against the baseline model for the bar chart) are non-significant. In other words, even though the waterfall plot adds information, especially with respect to additivity, it does not improve performance compared to the bar chart. We do observe a few significant effects in the second scenario of waterfall for pos/neg ($\beta_{waterfall:pos/neg:second} = \mathbf{1.37}, \boldsymbol{p < .01}$) and additivity questions ($\beta_{waterfall:addivity:second} = \mathbf{1.37}, \boldsymbol{p < .01}$). Together with the negative pos/neg and additivity effects for the second scenario ($\beta_{pos/neg:second} = \mathbf{-.96}, \boldsymbol{p < .01}$ and $\beta_{pos/neg:second} = \mathbf{-.96}, \boldsymbol{p < .01}$) these reflect that the waterfall seems do slightly better than the bar chart for these two questions types, as can also be seen in Figure 4, again reflecting a small learning effect in the second scenario.

Small differences in response times. We ran similar multilevel linear regressions on response time and found only small differences. Figure 5 shows that, in general, responses were slower for the additivity questions, which was supported by the strong effect of the additivity term in the model. Moreover, the model showed a significantly longer time for basic question times for the force plot, but not for the two other question types.

Personal traits play no major role (RQ2)

We measured visual and cognitive skills, as well as AI knowledge, at the start of the study. We included these skills in the multilevel regressions but did not find strong, significant effects of any of them. Visual skills and AI knowledge did not affect correctness, either

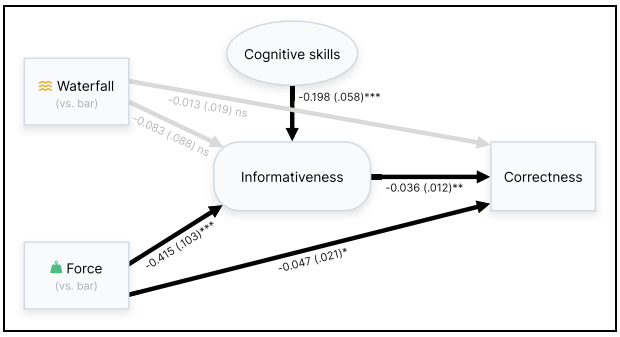


Figure 6. Structural equation model showing how the visualizations affect informativeness and total correctness score, in relation to personal characteristics (cognitive skills).
Standard errors are indicated in-between brackets, and significance with: $^{***}p<0.001$; $^{**}p<0.01$; $^{*}p<0.05$.

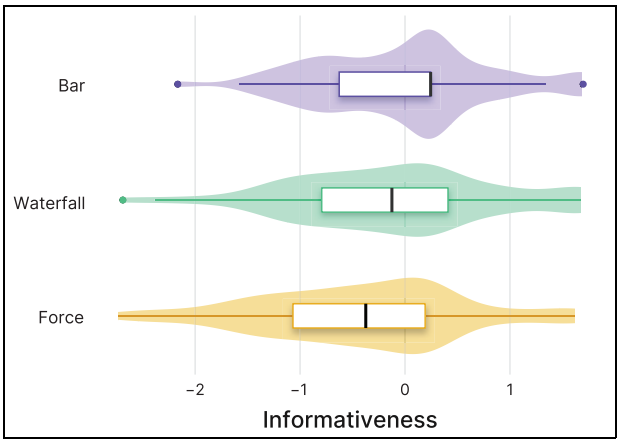


Figure 7. Distribution of the estimated factor score for perceived informativeness (based on Likert-scale questions) for each visualization type.

overall or for any of the individual visualizations. Cognitive skills did show some small but inconsistent interactions with some visualization or question types. As these effects were small and not consistent, we do not report on these in more detail here. We did not find any evidence that more complex visualizations (e.g. waterfall plot) resonated better with participants with higher visual or cognitive skills.

We expected personal characteristics to affect which visualizations are more effective (e.g. more expert users preferring more complex visualizations), but we found surprisingly little difference between visualizations. This means that our recommendations apply to all users regardless of visual literacy, cognitive skill, or experience with AI.

Subjective experiences (RQ3)

After showing each visualization, we also measured users' experience in terms of the perceived informativeness and effectiveness. To analyze how these relate to personal characteristics and to the number of correct responses, we ran a Confirmatory Factor Analysis (CFA) to test how the separate questions combine into their underlying factors. The CFA showed good convergent validity of all factors, personal characteristics, as well as informativeness and effectiveness. However, we found that informativeness and effectiveness were very highly correlated ($r>0.9$), thus lacking discriminant validity. As we are primarily interested in how informative the visualizations were, we continued with the informativeness measure and dropped the effectiveness measure. The CFA showed excellent fit after excluding Q5 of the informativeness measure and Q6 of the cognitive skill measure.

We then related informativeness to visualization conditions, the total proportion of correct answers, and the three personal characteristics in a path model. Starting with a fully saturated model, we found that only cognitive skills related significantly to informativeness, and visual skills and AI experience did not relate to any other factors. Excluding these other personal characteristics, the resulting path model (see Figure 6) showed an adequate fit: $\chi^2(50) = 93.855$, $p = 0.0002$, $RMSEA = 0.039$ $(90\%CI = [0.027, 0.051])$, $CFI = 0.993$, $TLI = 0.990$.

As in the earlier analyses, we chose the bar chart visualization as the baseline: the effects of the waterfall and force plots are compared against the bar chart. The path model shows that compared to the bar chart, the force plot is indeed perceived to be less informative, whereas the waterfall plot is not significantly different. The violin plots of the factor scores of informativeness (see Figure 7) indeed confirm that the force plot receives lower scores. Although the bar charts also seem to get somewhat higher and less negative responses than waterfall plots, these differences were not significant in our path model $(\beta = -0.083, p = 0.349)$. Higher cognitive skills positively affect perceived informativeness; in other words, people who score higher on cognitive skills perceive any visualization as more informative (We also investigated potential moderation of cognitive skills on the effect of visualization type on informativeness (e.g. do people with high cognitive skills find a more complex visualization more informative than people with a smaller reported cognitive skill?). We did not find any evidence of such moderation in our data.).

Most interesting is how the visualizations and informativeness relate to the proportion of total correct

answers. We find that higher perceived informativeness is related to increased correctness scores, showing that if people perceive a visualization as more informative, they also tend to score better. The waterfall plot does not impact the correctness score (relative to the bar chart), but the force plot shows lower correctness scores, as in our earlier objective performance analysis. However, this analysis shows that this is the result of two separate pathways: a direct effect of force plot on correctness ($\beta = -0.047, p = .024$) and an indirect effect mediated by informativeness ($\beta = -0.015, p = 0.010$; The total negative effect of force plot on correctness is significant ($\beta = -0.062, p = .003$).). In other words, part of the lower correctness of the force plot is explained by the force plot being less informative than the bar chart.

Follow-up study

After studying these newly introduced feature importance visualizations, we considered the design choices made for the visualizations we studied. For the first study, we stayed as close to the designs provided in the SHAP library to test a realistic use case. However, certain design decisions in those visualizations seem a bit counterintuitive. For example, the use of red for positive and blue for negative is opposite from the commonly used divergent color maps from ColorBrewer[58] and may cause mistakes. Next, waterfall plots are read from bottom to top to follow the addition from base rate to prediction, which may be counterintuitive (uncommon for visualizations and for many cultures in general). Finally, force plots place the positively contributing features on the left, whereas in all other visualizations, they are displayed on the right. This may be circumvented by changing the visual metaphor from "pushing" to "pulling" against the base rate.

The first goal of this follow-up study was to investigate whether these simple, yet logical design choices impacted participant performance. To this end, we conducted a follow-up study with two alternative visualizations to the waterfall and force plots, as shown in Figure 8. A second goal of the follow-up study was to see if we replicate the results and if with more data, we might be able to check for individual differences. The study was otherwise identical to the previous study, with the same baseline, setup, and questions.

We invited 300 more participants through Prolific, matching similar demographics to the first study. Of these, 279 passed the attention check and were analyzed for the follow-up study. More details about the participant sample are shown in Table 3; as in the original study, the full, raw demographic data collected from participants are available in the Supplemental material.

Main findings

Initial analyses reveal that results for the alternative designs are very similar to the original study, showing similar differences between the tested visualization types. The redesigned force and waterfall plots do not result in notable improvements in correctness, neither at the per-question level nor across the first and second visualizations (see Figure 9).

To test for any differences, we ran a similar multilevel logistic regression to predict correctness, combining the data of the two studies, and adding a predictor variable that indicated whether this was for the follow-up study with the alternative design or not. The model only shows a few additional effects for the design alternative. First, a negative main effect of the design alternative ($\beta_{Alternative} = -.46, p < .05$) showed that across the board, people did a bit worse, which is mostly reflected in the lower scores for the basic questions, as can be seen in Figure 10. We also observe an effect of ($\beta_{pos/neg:Alternative} = 0.52, p < .05$), which registers that the positive/negative questions were answered a bit better (relative to the basic questions) than in the original study. The figure also suggests that the force plot is not performing much worse, different from the original study. The evidence for this in the regression model is, however, weak; the effect of ($\beta_{force:Alternative} = 0.49, p < .1$) is positive but not significant, and there are no other significant interactions with the force plot. Apart from these effects, we find two (small) interactions with order, showing that the additive questions do better in the second scenario ($\beta_{additivity:second:Alternative} = 1.3, p < .05$). However, this does not hold true for the force plot, as the effect is counteracted by ($\beta_{force:additivity:second:Alternative} = -1.59, p < .05$). Overall, the design alternative shows only small differences from the original study, indicating that the alternative designs did not result in significantly better (or worse) performance. This reinforces the result from the first study that the new feature importance visualizations do not seem to help users understand aspects of feature performance any better than standard bar chart visualizations.

The role of individual differences

Given that the alternative design choices show very few differences compared to the original study, we can combine the two studies, which allows us to have more power to check the effect of visual and cognitive skills and AI knowledge. We first redid the structural equation model analysis as described in the previous subsection and found very similar results. Different from the original study, we now also find a small but

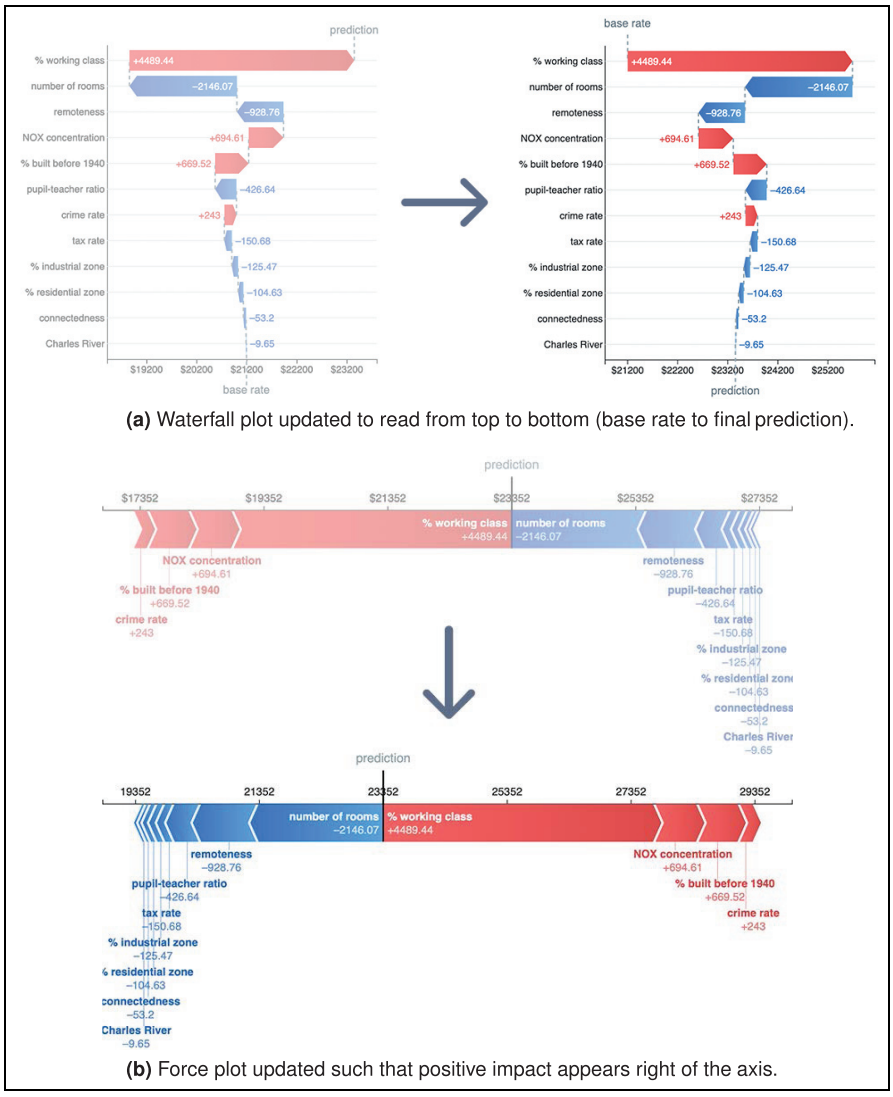


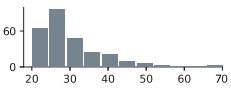
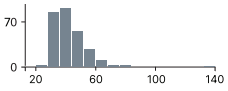
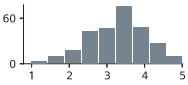
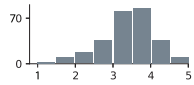
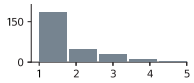
Figure 8. Updated visualizations with alternative designs: (a) Waterfall plot updated to read from top to bottom (base rate to final prediction), and (b) force plot updated such that positive impact appears right of the axis.
Note. The figure is shown for illustrative purposes only and is not intended for reading individual features or their values.

interesting effect of AI knowledge: whereas AI knowledge positively influences informativeness, it also negatively influences total correctness, showing that those having more AI knowledge perform slightly worse on the tasks. These two effects have opposite pathways in terms of the influence of AI knowledge on correctness. An analysis of the direct and indirect effect shows that the total indirect effect is small but positive ($\beta = 0.007, p < .05$), but the direct effect of knowledge on correctness is much stronger ($\beta = -.029$), resulting in a total negative effect ($\beta = -0.023, p < .05$). In other words, more AI knowledge improves informativeness, which increases correctness somewhat, but overall, AI knowledge seems to result in lower performance.

We further analyzed the role of individual differences on performance for the combined data, and as in the original study, we only found small effects. Both visual skills and AI knowledge did not significantly moderate any of the relations: that is, it was not the case that people with higher visual skills were better at understanding particular types of visualizations, or did better on specific types of questions. Given that we have a large sample size of 546 participants, this shows that these (self-reported) individual differences do not have substantial effects on how people understand and interpret the visualizations.

Only for cognitive skills, we found some small moderation effects. We ran a multilevel regression

Table 3. Summary of the participant demographics for the follow-up study.

Type	Answers (count)
Sex	Female **(157)**, Male **(121)**, Prefer not to say **(1)**
Nationality (47 total)	Europe **(153)**, Africa **(85)**, North America **(17)**, South America **(11)**, Asia **(10)**, Australia **(2)**
Age (years)	
Time taken (minutes)	
Cognitive skills	
Visualization familiarity	
Experience with (X)AI	

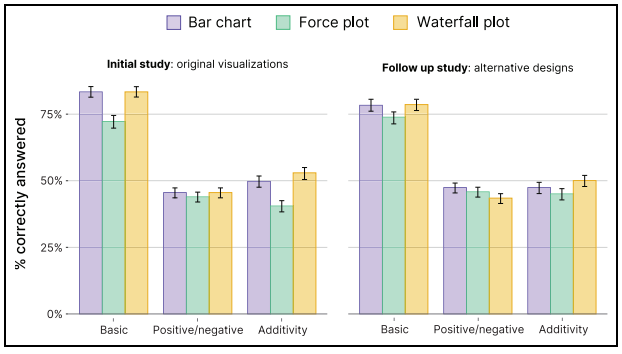


Figure 9. Correctness for each visualization across the three types of questions, for the original and alternative designs.

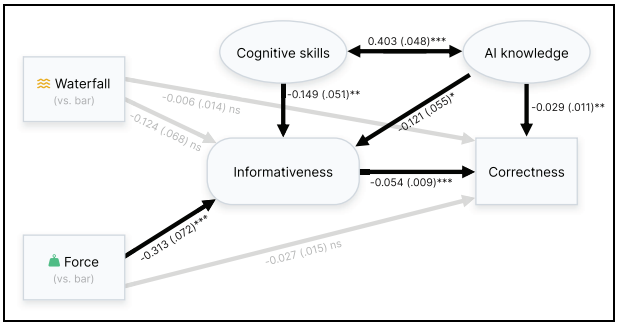


Figure 10. Updated structural equation model combining the first/original and follow-up studies.

model as before on correctness, with cognitive skill, visualization type, and question type as predictors (and their interactions). This model attempts to model the differences we can observe in Figure 11. Overall we do not observe a main effect of cognitive skills ($\beta_{cogHigh} = -0.32, p = .16$), but we observe an interaction of cognitive skill with question type additivity ($\beta_{cogHigh:additive} = 0.60, p < .01$), counteracted by another interaction with waterfall visualization ($\beta_{cogHigh:waterfall:additive} = -.76, p < .05$) and a non-significant but directionally similar effect for the force visualization ($\beta_{cogHigh:force:additive} = -.60, p = 0.075$). The figure also shows a somewhat better performance for additivity questions for high compared to low skills in the bar graph, whereas this effect is weaker for force and even turns around for the waterfall plots. This effect mostly shows that better cognitive skills do not necessarily lead to better performance for some visualizations (and sometimes even to lower).

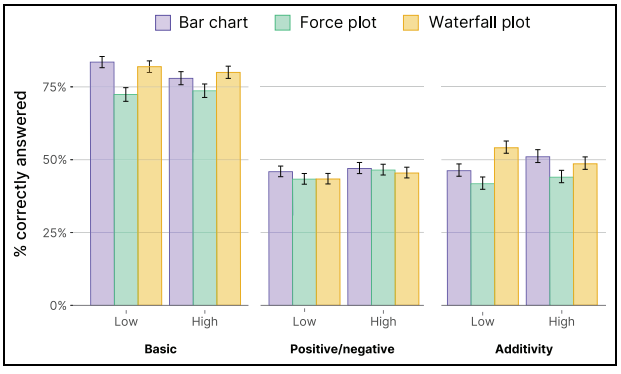


Figure 11. Correctness for each visualization across the three types of questions, comparing low and high cognitive skills per condition.

Discussion

In this work, we have performed an extensive user evaluation of commonly used SHAP visualizations, measuring both the (objective) performance of users'

understanding of the plots, as well as their (subjective) perceptions. Our large sample size allows us to test for the potentially small differences and to do adequate tests of individual differences, such as cognitive skills, visual familiarity, and AI knowledge.

As we noted in the previous section, the main difference between the three visualizations is that participants perform worse using the force plot, and the bar chart and waterfall plot lead to similar performances. This is an interesting observation, as these two plot types convey different information. The bar chart only reflects the values directly, whereas the waterfall plot includes the aspect of additivity, where the visual representation emphasizes that importance scores can be combined to obtain the prediction. As such, one could argue that the waterfall plot should be more helpful in answering additivity questions. One explanation for the lack of improved performance using a waterfall plot could be its associated cognitive load, where the positive effect of the added data is negated by the load associated with this added information.

In our first experiment, we intentionally evaluated the default SHAP visualizations, as these represent what users most commonly encounter in practice. In the follow-up study, we also tested modified versions of the force and waterfall plots to reduce design-specific factors; however, these variations did not lead to measurable performance improvements. The three visualizations differ not only in their metaphor but also in low-level design properties, such as orientation and spatial layout. In particular, the force plot uses a horizontal arrangement, whereas the bar and waterfall plots use a vertical layout. This may have introduced a confounding factor, as participants might have been more familiar with vertically oriented bar-like charts or found horizontal encodings harder to parse. Therefore, the observed performance differences between the force plot and the other two conditions may partially reflect layout effects rather than the visualization metaphor alone.

We also noted that the correctness of Q6 and Q8 was especially low compared to the rest of the questions. Q5 and Q6 ("*What happens to the predicted house price if feature X* increases?") were meant to test the positive/negative property, but they also relate to the *actionability* property: can you infer what would happen to the prediction if you change the feature value? We find that 91% of participants expected the prediction to increase or decrease based on the feature importance. However, you cannot infer this from (Shapley-based) feature importance alone. We expect participants may have inferred what would happen through common sense (e.g. if the crime rate goes up, surely it negatively affects the housing price). This is a big problem as it conflates the expected causal relationship in the real world with the correlational inner-workings of the model. These results match with prior work that showed that users may expect actionability from a feature importance method even if it does not exhibit this property.[5]

As for Q8, there were many more answer options (multiple selections out of all features) compared to other questions (four multiple-choice answers), which may have affected the absolute correctness score. We also asked for the "minimal set" of features that puts the house price above a certain value to ensure only one correct answer exists, but this may have been missed by many of our participants. We checked all answers again manually and found that 52% of participants would have been correct if we had neglected the "minimal set" requirement from the question, comparable to the other questions of the same type. Note that we also ran our analyses excluding Q6 and Q8, but did not find any substantial differences in our model results that would affect our main findings and interpretations.

Limitations

To be able to evaluate these visualizations in isolation, we made choices in the study design that have implications. First, we tested only one type of regression task, but feature importance can also be used for classification. While classification involves probabilities or discrete labels rather than continuous values, we expect the interpretation of the tested visualizations to remain largely similar. Therefore, we do not expect this to affect the results much, but we cannot rule it out either.

Furthermore, we only tested one type of scenario, namely predicting housing prices. This provided context makes the questions more manageable for novices, but may affect our results and generalizability. However, the tested visualizations and quiz tasks focus on general properties of feature importance explanations (ranking, sign of contribution, and additivity), which are not specific to housing data and commonly appear across domains.

Next, in our study, we assumed that feature importance provides accurate explanations of the model. Although feature importance is widely used and a trusted explanation technique, explanation techniques can fail or have low fidelity to the reference model. The visualizations we evaluate in this paper do not incorporate the uncertainty associated with the technique, and as such, may lead to unexpected or incorrect conclusions about the model.

Lastly, we tested for our participants' understanding of feature importance as a proxy for the quality of their understanding of the model itself. If users do not fully

understand the properties of the explanation, they may be misled or draw incorrect conclusions. However, this is only part of what makes a good explanation, and as such, there are several avenues for future research.

Future work

This paper has provided a quantitative look at the comparison between feature importance visualizations. A natural extension of our work would be to look closely into how people read and interpret plots through qualitative methods. Using techniques such as eye tracking and think-aloud protocols, we could gain deeper insights into how visualizations are read and interpreted, and might shed more light on the similarity of the results we observed between bar charts and waterfall plots.

The visualizations in this study concerned the explanation of model predictions without conveying any kind of uncertainty. Feature importance explanations from SHAP are approximations and may not always represent the model faithfully. None of the visualizations encompasses such information, and including this uncertainty is an interesting direction for visualization research.

In this work, we have exclusively focused on visualizations for *local* explanations concerning a single prediction. Based on our power analysis, extending the study to include additional visualization types would not have been feasible without substantially increasing the required sample size. As a result, we deliberately limited the scope of the study. Future work could extend our methodology to global visualizations (e.g. beeswarm plots, decision plots, and heatmaps) provided by the SHAP library. Future work could also replicate our study across different application domains (e.g. healthcare or finance) and prediction tasks (e.g. classification) to further assess the generalizability of our findings.

Finally, while our study focuses on SHAP, similar trade-offs between expressiveness and cognitive simplicity arise in other XAI methods that communicate feature attributions or importance. We hypothesize our results generalize to other feature importance techniques (e.g. LIME, GAM), but future work could more rigorously investigate the differences between, and applicability to explanation approaches. Moreover, hybrid visualization designs may be promising, for example: combining the simplicity of bar charts with optional additivity cues through progressive disclosure. Such designs may be particularly beneficial in scenarios where users need to understand how a prediction is constructed rather than only compare feature rankings.

Conclusion

We analyzed different visualizations used to convey feature importance to users, including standard bar charts, but also advanced visual metaphors such as waterfall plots and force plots. These advanced visual representations convey more information (e.g. about the additivity property of the technique), but the added benefit from these encodings is poorly understood and has not yet been formally studied.

In this work, we tested specifically for the aspects that advanced visualizations added over standard bar chart visualizations through 2 user experiments with 546 participants. We studied both subjective and objective measures, but contrary to our initial hypotheses, found no evidence that these visualizations helped users understand aspects of feature performance any better; not in terms of objectively measured performance on the questions, nor completion time, nor subjective informativeness. In fact, the force plot was consistently worse in our testing. At least for local explanations, our findings lead us to recommend avoiding force plot visualizations for feature importance.

While bar and waterfall plots showed little difference in performance, we found no evidence that the arguably more complex waterfall visualization provides any benefit. We also experimented with waterfall and force plots with a modified design. We hypothesized that these may improve performance in the study, but our results did not reveal a difference in performance either. Hence, we would tentatively suggest using bar charts to convey feature importance.

We did expect personal characteristics to affect which visualizations are more effective (e.g. more expert users preferring more complex visualizations), but we found surprisingly little difference between the studied visualization types. This means that our recommendations apply to all users regardless of visual literacy, cognitive skill, or experience with AI.

ORCID iDs

Dennis Collaris https://orcid.org/0000-0001-7612-9319
Yu Liang https://orcid.org/0000-0002-4034-2211
Martijn C. Willemsen https://orcid.org/0000-0001-5908-9511
Angelos Chatzimparmpas https://orcid.org/0000-0002-9079-2376
Jarke J. van Wijk https://orcid.org/0000-0002-5128-976X

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or

publication of this article: This work is part of the research programme TEPAIV with project number 612.001.752, which is financed by the Dutch Research Council (NWO), and has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101206814.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

References

1. Harrison D and Rubinfeld DL. Hedonic housing prices and the demand for clean air. *J Environ Econ Manag* 1978; 5(1): 81–102.
2. Lundberg SM and Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.) *Proceedings of the advances in neural information processing systems*, Vol. 30. 2017, Curran Associates, Inc.
3. Dimara E and Stasko J. A critical reflection on visualization research: where do decision making tasks hide? *IEEE Trans Vis Comput Graph* 2022; 28(1): 1128–1138.
4. Collaris D and van Wijk JJ. Explain explore: Visual exploration of machine learning explanations. In: *Proceedings of the 2020 IEEE Pacific visualization symposium (PacificVis)*, 2020, pp. 26–35. IEEE.
5. Collaris D, Weerts HJ, Miedema D, et al. Characterizing data scientists' mental models of local feature importance. In: *Proceedings of the Nordic human-computer interaction conference, NordiCHI '22*, 2022. Association for Computing Machinery, ISBN 9781450396998.
6. Isenberg T, Isenberg P, Chen J, et al. A systematic review on the practice of evaluating visualization. *IEEE Trans Vis Comput Graph* 2013; 19(12): 2818–2827.
7. Kosara R. An empire built on sand: reexamining what we think we know about visualization. In: *Proceedings of the sixth workshop on beyond time and errors on novel evaluation methods for visualization*, 2016, pp. 162–168.
8. Chatzimparmpas A, Kucher K and Kerren A. Visualization for trust in machine learning revisited: the state of the field in 2023. *IEEE Comput Graph Appl* 2024; 44: 99–113.
9. Chatzimparmpas A, Martins RM, Jusufi I, et al. A survey of surveys on the use of visualization for interpreting machine learning models. *Inf Vis* 2020; 19(3): 207–233.
10. Chatzimparmpas A, Martins RM, Jusufi I, et al. The state of the art in enhancing trust in machine learning models with the use of visualizations. *Comput Graph Forum* 2020; 39: 713–756. https://doi.org/10.1111/cgf.14034
11. Liu S, Wang X, Liu M, et al. Towards better analysis of machine learning models: a visual analytics perspective. *Vis Inform* 2017; 1(1): 48–56.
12. Molnar C (ed). *Interpretable machine learning*. 3rd ed. 2025. ISBN 978-3-911578-03-5.
13. Cashman D, Humayoun SR, Heimerl F, et al. A user-based visual analytics workflow for exploratory model analysis. *Comput Graph Forum* 2019; 38(3): 185–199.
14. Yuan J, Chan GY, Barr B, et al. SUBPLEX: a visual analytics approach to understand local model explanations at the subpopulation level. *IEEE Comput Graph Appl* 2022; 42(6). 24–36. https://doi.org/10.1109/MCG.2022.3199727
15. Wang Q, Huang K, Chandak P, et al. Extending the nested model for user-centric XAI: a design study on GNN-based drug repurposing. *IEEE Trans Vis Comput Graph* 2023; 29(1). 1266–1276. https://doi.org/10.1109/TVCG.2022.3209435
16. Angelini M, Blasilli G, Lenti S, et al. A visual analytics conceptual framework for explorable and steerable partial dependence analysis. *IEEE Trans Vis Comput Graph* 2024; 30(8). 4497–4513. https://doi.org/10.1109/TVCG.2023.3263739
17. Kerrigan D, Barr B and Bertini E. PDPilot: exploring partial dependence plots through ranking, filtering, and clustering. *IEEE Trans Vis Comput Graph* 2025; 31(10). 7377–7390. https://doi.org/10.1109/TVCG.2025.3545025
18. Casalicchio G, Molnar C and Bischl B. Visualizing the feature importance for black box models. In: *Proceedings of the machine learning and knowledge discovery in databases*, 2019, pp. 655–670. Springer International Publishing.
19. Kumar P and Sharma M. Feature-importance feature-interactions (FIFI) graph: a graph-based novel visualization for interpretable machine learning. In: *Proceedings of the 2021 international conference on intelligent technologies (CONIT)*, 2021, pp. 1–7. https://doi.org/10.1109/CONIT51480.2021.9498467.
20. Munzner T. Visualization analysis and design. In: *Proceedings of the special interest group on computer graphics and interactive techniques conference courses, SIGGRAPH Courses '25*, 2025. Association for Computing Machinery, ISBN 9798400715433.
21. Cleveland WS and McGill R. Graphical perception: theory, experimentation, and application to the development of graphical methods. *J Am Stat Assoc* 1984; 79(387): 531–554.
22. Heer J and Bostock M. Crowdsourcing graphical perception: using mechanical Turk to assess visualization design. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, CHI '10, 2010, p. 203–212. Association for Computing Machinery. ISBN 9781605589299.
23. Ribeiro MT, Singh S and Guestrin C. "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, 2016, pp. 1135–1144.
24. Hadash S, Willemsen MC, Snijders C, et al. Improving understandability of feature contributions in model-agnostic explainable AI tools. In: *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1–9.
25. Few S (ed). *Show me the numbers: designing tables and graphs to enlighten*. 2nd ed. Analytics Press, 2012. ISBN 0970601972.
26. Correll M. Ethical dimensions of visualization research. In: *Proceedings of the 2019 CHI conference on human factors in computing systems, CHI'19*, 2019, p. 1–13. Association for Computing Machinery, ISBN 9781450359702.
27. Shapley LS. A value for n-person games. In: Kuhn HW (ed.) *Contributions to the theory of games*. Princeton University Press, 1953, Vol. 2, pp. 307–318.
28. Kononenko I. An efficient explanation of individual classifications using game theory. *J Mach Learn Res* 2010; 11: 1–18.
29. Huang M, Chen C and Sun LZ. Statistical interpretation and comparison of waterfall plots. *JCO Clinical Cancer Informatics* 2023; 7(7). e2300132. https://doi.org/10.1200/CCI.23.00132
30. Gillespie TW. Understanding waterfall plots. *J Adv Pract Oncol* 2012; 3(2): 106–111.
31. Hohman F, Head A, Caruana R, et al. 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*, 2019, pp. 1–13.
32. Nagy M and Molontay R. Interpretable dropout prediction: towards XAI-based personalized intervention. *Int J Artif Intell Educ* 2024; 34: 274–300. https://doi.org/10.1007/s40593-023-00331-8
33. Liu Y, Song C, Tian Z, et al. Identification of high-risk patients for postoperative myocardial injury after CME using machine learning: a 10-year multicenter retrospective study. *Int J Gen Med* 2023; 16: 1251–1264. https://doi.org/10.2147/IJGM.S409363
34. Stenwig E, Salvi G, Rossi PS, et al. Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Med Res Methodol* 2022; 22(1). 53. https://doi.org/10.1186/s12874-022-01540-w
35. Moreira Cunha B and Diniz Junqueira Barbosa S. Evaluating the effectiveness of visual representations of SHAP values toward explainable artificial intelligence. In: *Proceedings of the XXIII Brazilian symposium on human factors in computing systems, IHC '24*, 2024. Association for Computing Machinery, ISBN 9798400712241.
36. Al-Ansari N, Al-Thani D and Al-Mansoori RS. User-centered evaluation of explainable artificial intelligence (XAI): a systematic literature review. *Hum Behav Emerg Technol* 2024; 2024(1): 4628855.
37. Paleja R, Ghuy M, Arachchige NR, et al. The utility of explainable AI in ad hoc human-machine teaming. In: *Proceedings of the 35th international conference on neural information processing systems, NIPS '21*, 2021. Curran Associates Inc, ISBN 9781713845393.
38. Parsa AB, Movahedi A, Taghipour H, et al. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accid Anal Prev* 2020; 136: 105405.
39. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018; 2(10): 749–760.
40. Rodríguez-Pérez R and Bajorath J. Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions. *J Comput Mol Des* 2020; 34(10): 1013–1026.
41. Kaur H, Nori H, Jenkins S, et al. 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*, 2020, pp. 1–14.
42. Conati C, Barral O, Putnam V, et al. Toward personalized XAI: a case study in intelligent tutoring systems. *Artif Intell* 2021; 298: 103503.
43. Ghai B, Liao QV, Zhang Y, et al. Explainable active learning (XAL): toward AI explanations as interfaces for machine teachers. *ACM Hum-Comput Interact* 2021; 4(3): 1–28.
44. Millecamp M, Htun NN, Conati C, et al. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In: *Proceedings of the 24th international conference on intelligent user interfaces, IUI '19*, 2019, p. 397–407. Association for Computing Machinery, ISBN 9781450362726.
45. Chatti MA, Guesmi M, Vorgerd L, et al. Is more always better? The effects of personal characteristics and level of detail on the perception of explanations in a recommender system. In: *Proceedings of the 30th ACM conference on user modeling, adaptation and personalization, UMAP '22*, 2022, p. 254–264. Association for Computing Machinery, ISBN 9781450392075
46. Cacioppo JT and Petty RE. The need for cognition. *J Pers Soc Psychol* 1982; 42(1): 116–131.
47. Yeon J and Rahnev D. The suboptimality of perceptual decision making with multiple alternatives. *Nat Commun* 2020; 11: 3857. https://doi.org/10.1038/s41467-020-17661-z
48. Gleicher M. Considerations for visualizing comparison. *IEEE Trans Vis Comput Graph* 2018; 24(1). 413–423. https://doi.org/10.1109/TVCG.2017.2744199
49. Kouki P, Schaffer J, Pujara J, et al. Personalized explanations for hybrid recommender systems. In: *Proceedings of the 24th international conference on intelligent user interfaces*, IUI '19, 2019, p. 379–390. Association for Computing Machinery, ISBN 9781450362726.
50. Lins de Holanda Coelho G, Hanel PH and Wolf JL. The very efficient assessment of need for cognition: developing a six-item version. *Assess* 2020; 27(8): 1870–1885.

51. Silva A, Schrum M, Hedlund-Botti E, et al. Explainable artificial intelligence: evaluating the objective and subjective impacts of XAI on human-agent interaction. *Int J Hum Interact* 2023; 39(7): 1390–1404.
52. Hoffman RR, Mueller ST, Klein G, et al. Metrics for explainable AI: challenges and prospects. CoRR 2018; abs/1812.04608: 1–50.
53. Adhikari A, Tax DMJ, Satta R, et al. LEAFAGE: example-based and feature importance-based explanations for black-box ml models. In: *Proceedings of the 2019 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, 2019, pp. 1–7.
54. Liang Y and Willemsen MC. Interactive music genre exploration with visualization and mood control. In: *Proceedings of the 26th international conference on intelligent user interfaces*, IUI '21, 2021, pp. 175–185. Association for Computing Machinery, ISBN 9781450380171.
55. Holzinger A, Carrington A and Müller H. Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations. *Künstl Intell* 2020; 34(2): 193–198.
56. Knijnenburg BP, Willemsen MC, Gantner Z, et al. Explaining the user experience of recommender systems. *User Model User-Adapt Interact* 2012; 22: 441–504.
57. Brooke J. SUS: a quick and dirty usability scale. *Usability Eval Ind* 1996; 189(3): 189–194.
58. Harrower M and Brewer CA. Colorbrewer.org: an online tool for selecting colour schemes for maps. *Cartogr J* 2003; 40(1): 27–37.

Appendices

Questions for individual differences

Visual familiarity perceived informativeness (five-point Likert scale)

- I am competent when it comes to graphing and tabulating data.
- I frequently tabulate data with computer software.
- I have graphed a lot of data in the past.
- I frequently analyze data visualizations.

Cognitive skills perceived informativeness (five-point Likert scale)

- I would prefer complex to simple problems.
- I like to have the responsibility of handling a situation that requires a lot of thinking.
- Thinking is not my idea of fun. *(Reversal)*
- I would rather do something that requires little thought than something that is sure to challenge my thinking abilities. *(Reversal)*
- I really enjoy a task that involves coming up with new solutions to problems.
- I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought

Experience with AI and XAI

- How much do you know about artificial Intelligence or machine learning algorithms? Choices:
  - I have no knowledge of AI or ML algorithms
  - I have heard of AI or ML algorithms
  - I know basic concepts in AI or ML algorithms
  - I have used AI algorithms or ML algorithms before
  - I have demonstrated expertise in AI or ML algorithms
- How much do you know about regression algorithms? Choices:
  - I have no knowledge of regression algorithms
  - I have heard of regression algorithms
  - I know basic concepts of regression algorithms
  - I have used regression algorithms before
  - I have demonstrated expertise in regression algorithms
- How much do you know about Explainable AI? Choices:
  - I have no knowledge of it
  - I have heard of it
  - I know basic concepts of it
  - I have used it before
  - I have demonstrated expertise in it
- If applicable, briefly describe the Explainable AI techniques you have heard of or used. *(open question)*.

Questions for objective measures

The questions in this section were randomized to prevent learning effects.

Basic understanding

Q1. Which of the following statements is correct? (*Multiple-choice, four options following the template: "The feature X is the most important contributor to the predicted house price."*)

Q2. Which of the following statements is correct? (*Multiple-choice, four options following the template:*

"The feature X is among the top three [most/least] important contributors to the predicted house price.")

Positive/negative impact

Q3. Which of the following statements is correct? (select all that apply; *Multiple-choice with two correct answers, four options following the template: "The feature X has a positive / negative impact on the predicted house price."*)

Q4. Which of the following statements is correct? (*Multiple-choice, four options following the template: "The feature X is the* first / second / third *most important positive / negative contributor to the predicted house price."*)

Q5. Which features would increase the predicted housing price if they were not considered by the model? (Select all that apply; *Multiple choice out of four random features.*)

Q6. What happens to the predicted house price if the feature X increases? Choices:

- The predicted housing price will increase
- The predicted housing price will decrease
- The predicted housing price will stay the same
- You cannot tell

Additivity

Q7. Features X and Y have a combined effect on the house price of: (*Multiple choice of four different price ranges.*)

Q8. Select a minimal set of features that, if considered by the model, would put the housing price above $ price: (*Multiple-choice, check boxes for all features in the dataset.*)

Q9. Select a feature combination so that they put housing price above $ price. (*Multiple-choice, check boxes for four features in the dataset.*)

Questions for subjective measures

Perceived informativeness (five-points Likert scale)

- The visual explanation was informative.
- This visual explanation had sufficient detail of how the prediction model works.
- The visual explanations provided enough information for me to understand how the prediction model works.
- The visual explanation provided me with sufficient information to answer the questions.
- The visual explanations were not relevant for the questions I was given. *(Reversal)*

Perceived effectiveness (five-points Likert scale)

- I could easily follow the visual explanation to arrive at an answer to the question.
- The visual explanations were useless. *(Reversal)*
- The explanations were actionable, that is, they helped me know how to answer the questions.
- I could better answer the questions with the help of visual explanations.