J Vis (2022) 25:47–57
https://doi.org/10.1007/s12650-021-00776-w

REGULAR PAPER

Dennis Collaris · **Jarke J. van Wijk**

Comparative evaluation of contribution-value plots for machine learning understanding

Received: 1 May 2021 / Revised: 30 June 2021 / Accepted: 30 July 2021 / Published online: 11 September 2021

Abstract The field of explainable artificial intelligence aims to help experts understand complex machine learning models. One key approach is to show the impact of a feature on the model prediction. This helps experts to verify and validate the predictions the model provides. However, many challenges remain open. For example, due to the subjective nature of interpretability, a strict definition of concepts such as the contribution of a feature remains elusive. Different techniques have varying underlying assumptions, which can cause inconsistent and conflicting views. In this work, we introduce local and global contribution-value plots as a novel approach to visualize feature impact on predictions and the relationship with feature value. We discuss design decisions and show an exemplary visual analytics implementation that provides new insights into the model. We conducted a user study and found the visualizations aid model interpretation by increasing correctness and confidence and reducing the time taken to obtain an insight.

Keywords Visualization · Machine Learning · Explainable AI · Interpretability

1 Introduction

The past decade has witnessed a sharp increase in the popularity of artificial intelligence and machine learning. This prevalence has resulted in a wide variety of new approaches and techniques (e.g., deep learning) that have achieved astounding results previously not deemed possible (McKinney et al. 2020; Karras et al. 2019). Clearly, these models have advanced over their predecessors in terms of predictive performance (e.g., accuracy, precision, recall, F1-score). However, there are more properties of these models that have not received as much attention, such as complexity, interpretability, and fairness (Doshi-Velez 2017). As a consequence, state-of-the-art techniques are ever increasing in complexity, yielding black-box models that cannot easily be inspected or verified.

The field of explainable artificial intelligence (XAI) has recently gained a lot of traction as it aims to alleviate these issues. It exposes more details about the behavior of complex machine learning models, which helps experts to verify and validate model predictions. XAI has proposed a variety of new techniques to show the impact of a feature on the model prediction (Friedman 2001; Goldstein et al. 2015; Ribeiro et al. 2016; Lundberg and Lee 2017). However, due to the novelty of the field many challenges remain open.

In particular, the complex and ill-defined nature of interpretability hinders a strict definition of concepts such as contribution of a feature. Different techniques have varying underlying assumptions, which can cause different and conflicting results. In this work, we present local and global contribution-value plots as a novel technique to explain machine learning models. The plots visualize the feature contribution to a

D. Collaris (✉) · J. J. van Wijk
Eindhoven University of Technology, Eindhoven, The Netherlands
E-mail: d.a.c.collaris@tue.nl

prediction, as well as the relationship with feature value. Such information about the model is typically conveyed with multiple techniques, which could lead to contradictory results. We discuss relevant design decisions and show an exemplary visual analytics instrumentation and show it enables insights into the model that were previously not possible.

To validate our proposed technique, we conducted a comparative user study with a variety of machine learning professionals and visualization experts. The results show that our visualizations aid model interpretation by increasing correctness and confidence and reducing the time taken to obtain an insight.

This article is an extension of a paper that was originally published at the international symposium on visual information communication and interaction (VINCI) (Collaris and van Wijk 2020b). In the original version, we did not report on the comparative user study just mentioned.

2 Background and related work

Visualization can help data scientists to get a better understanding of black box models. For trivial prediction problems, this can be done by inspecting the predictions of a model directly (Fig. 1a). Scatter plots can be used to show the relationship between prediction probability $\hat{y}$ and feature value $\mathbf{x}$. However, for any non-trivial prediction problems, there are likely many interactions between features which make it impossible to identify patterns and trends.

2.1 Local partial dependence plot

To help to gain insight into models, Friedman (2001) introduced the partial dependence plot (PDP). This is a sensitivity analysis technique that shows how the prediction $\hat{y}$ changes as the features of interest $\mathbf{z}_t$ (i.e., target features) are varied over their marginal distributions (Fig. 1b).

To define partial dependence for a data point $\mathbf{x}$, let $\mathbf{z}_t \subset \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of target features, and $\mathbf{z}_c$ the complement of $\mathbf{z}_t$ such that

$$\mathbf{z}_c \cup \mathbf{z}_t = \mathbf{x}, \quad \mathbf{z}_c \cap \mathbf{z}_t = \emptyset \tag{1}$$

The prediction $\hat{f}(\mathbf{x})$ in principle depends on both subsets:

$$\hat{y} = \hat{f}(\mathbf{x}) = \hat{f}(\mathbf{z}_t, \mathbf{z}_c) \tag{2}$$

However, if we fix the specific values of features in $\mathbf{z}_c$, then $\hat{f}(\mathbf{x})$ can be considered as a function only dependent on $\mathbf{z}_t$. This function represents the *local partial dependence* of the features in $\mathbf{z}_t$

$$\hat{f}_{\mathbf{z}_c}(\mathbf{z}_t) = \hat{f}(\mathbf{z}_t \mid \mathbf{z}_c) \tag{3}$$

If $\mathbf{z}_t$ consists of a single feature, a line graph of this function shows how changing $\mathbf{z}_t$ impacts the prediction of a single data point. This conveys much more about the model than just showing the prediction for single points and has been used in prior visualization work to explain machine learning (Krause et al. 2016a, b).

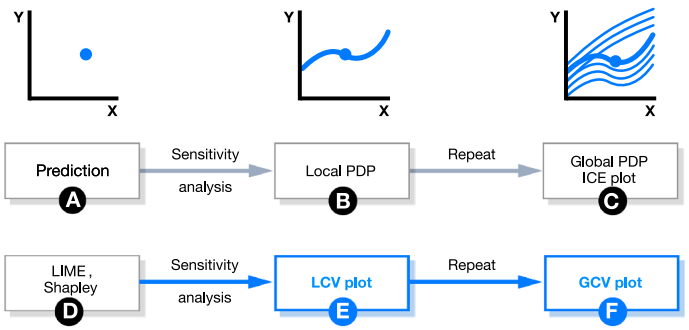


Fig. 1 Design space of interpretability methods. Blue boxes indicate our contribution. The *x*-axis denotes feature values; the *y*-axis denotes either prediction probability **a**, **b**, **c** or feature contribution **d**, **e**, **f**

2.2 Global PDP and ICE plot

Local PDPs provide a great insight into a single prediction. However, for many applications such a local explanation is not sufficient. In an explorative setting, experts would like to inspect much more than just a single prediction. For example, the explanation of a single prediction is not helpful for diagnosing problems with a model, or for model refinement. Even if there is a single prediction of interest, instance-level explanations do not show whether they are specific to that instance, or generalize to a larger set of instances. For these cases, we need *global* explanations. To get a global insight into the entire model, Friedman (2001) proposes averaging local partial dependence lines of all N training data points as follows:

$$\begin{aligned} \bar{f}_t(\mathbf{z}_t) &= \mathbb{E}_{\mathbf{z}_c}\left[\hat{f}(\mathbf{x})\right] \\ &= \int \hat{f}(\mathbf{z}_t, \mathbf{z}_c) M_c \, d\mathbf{z}_c \approx \frac{1}{N}\sum_{i=0}^{N} \hat{f}(\mathbf{z}_t, \mathbf{z}_c^N) \end{aligned} \tag{4}$$

where M_c is the marginal probability density of $\mathbf{z}_c$. This global PDP is used in visualization work to explain and compare machine learning models (Zhao et al. 2018; Wexler et al. 2019). However, Friedman notes that Eq. (4) does not hold when there is a strong interdependence amongst features, which is often the case for complex black box models.

To deal with interdependence, Goldstein et al. (2015) proposed an alternative called individual conditional expectation (ICE) plot by superimposing all individual local partial dependence lines. This reveals patterns that would otherwise be hidden by averaging. For example, the plot in Fig. 1C shows two clusters of partial dependence lines that would not be apparent in a global PDP.

2.3 Feature contribution

An alternative approach to gain insight into machine learning models is the feature contribution technique (Fig. 1d). Such methods yield feature contribution vectors that indicate how much every feature contributed to a prediction.

Initially, Baehrens et al. (2010) showed that machine learning models can be explained using the derivative of the class probability function. The reasoning is that if a small change in feature value leads to a large change in the prediction probability (or regression output), that feature is relevant for the prediction. They note, however, that an exact derivative for the majority of models does not exist.

To this end, LIME was proposed by Ribeiro et al. (2016). It solves this issue by fitting a linear regression surrogate model to the class probability gradient with a local sampling region around an instance. The coefficients of the linear model effectively approximate the derivative of the probability function, regardless of whether a formal derivative exists. Next, the approximation can be used to show which features have the most impact on a prediction.

Another prominent approach for feature contribution is Shapley values (Kononenko et al. 2010; Štrumbelj et al. 2009; Lundberg and Lee 2017). This method estimates the contribution of a feature by comparing the class probability of a prediction including and not including this feature (Merrick and Taly 2019). The absence of a feature is estimated by averaging the predictions for different values for that feature sampled from the training data distribution.

Any of these techniques yield feature contribution vectors that give a quick overview of which feature had an impact on a single prediction. However, it remains unclear for which *values* in general that feature is relevant. For example, in a medical trial where feature attribution shows that 'dosage' is important predictor for recovery, we would also like to know what values of 'dosage' were most relevant. In addition, these methods only target single predictions, whereas some use cases require a global perspective on the model.

Finally, various other explainable AI visualization works exist (Guidotti et al. 2018), but those often use the presented elementary techniques as a basis.

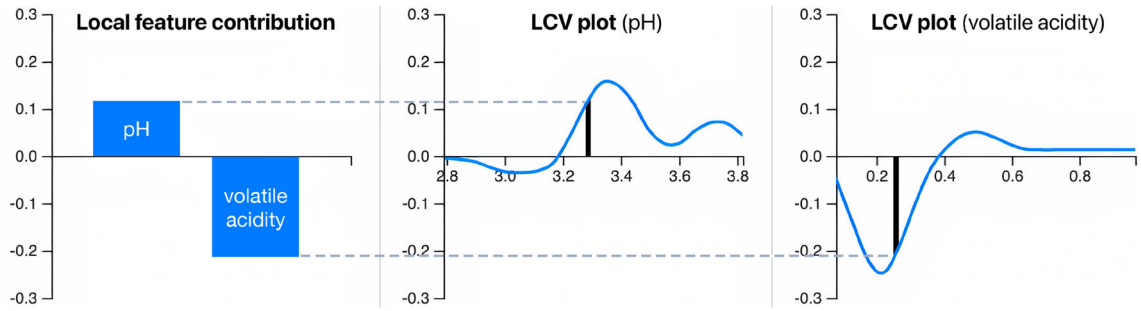


Fig. 2 LCVs can be directly compared with feature contribution visualizations. Left: contribution for two features of instance *i* represented as a bar chart. Right: LCV plots for the same features for full feature range. Value of instance *i* indicated with the vertical black line

3 Local contribution-value plot

To alleviate the limitations of previous techniques, we propose the local contribution-value (LCV) plot. The curves are generated in the same way as PDPs (Sect. 2.1), but instead of class probability values we use feature contribution values (Fig. 1E). This yields a plot that reveals how the *feature contribution* varies for changes in feature value. It has some key advantages over local PDPs.

First, contrary to a PDP (Friedman 2001), the LCV plot is also effective when features are heavily correlated. For example, if feature k and l are correlated, changing the value of either does not change the prediction, while changing both would. As the sensitivity analysis used in PDPs only alters the value of a single feature at a time, the PDP would not show variation in the prediction. In contrast, LCV plots use an explanation technique that considers a wider region of feature space (compared to a single point), which enables them to show variation in contribution even when features are correlated.

Next, for certain use cases the LCV plot may be easier to read and compare. To infer relevance of a feature in a PDP plot, experts have to consider the *slope* of the line. Previous work has shown that human slope estimation is not trivial and prone to be biased (i.e., angle contamination) as our visual system is geared towards judging angle rather than slope (Cleveland and McGill 1985). Hence, our graphical perception of slopes prohibits any exact judgment of contribution or importance in prediction-value plots. LCV plots encode feature contribution with position, making it easier to read and compare exact values (see Fig. 2).

This does mean that the prediction probability is not directly encoded in LCV plots. We argue PDP and LCV plots serve a different (and complementary) purpose. When experts are interested in the predictions for specific data points, PDPs are more suitable. However, when trying to understand how the model makes predictions, LCV plots are more suitable.

Finally, the LCV remains a local approach focusing on a single instance. This makes it difficult to get a global overview of the model and whether a feature that is locally relevant is always relevant, or only for a small number of instances.

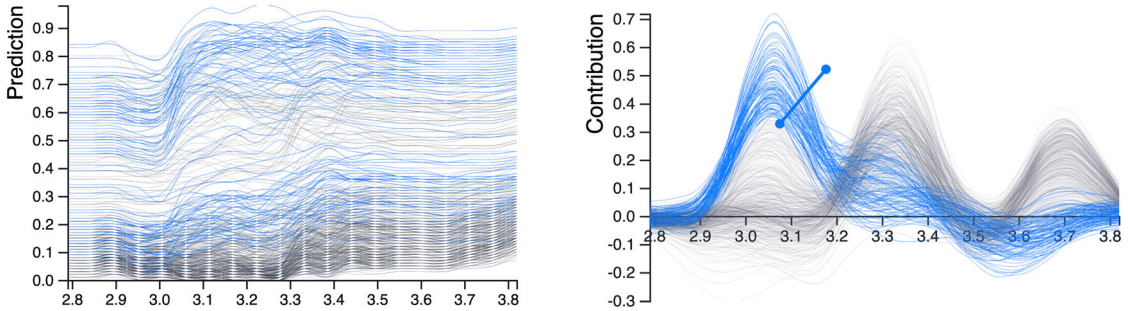
4 Global contribution-value plot

For a global overview, we propose using the same procedure as for an ICE plot: to superimpose LCV plots to show the contribution for an entire dataset (Fig. 1f). This helps experts to get a global overview of the model behavior for typical data. We refer to this approach as the global contribution-value (GCV) plot.

The GCV shows more clearly which values of a feature have a significant impact on the model prediction, which helps to understand the model. As an example, we examine the Wine Quality dataset (Cortez et al. 2009). Figure 3b shows two different thresholds (3.05 and 3.35) for pH that the random forest model uses to determine wine quality.

Next, GCV plots enable the comparison of feature importance at different feature values. For instance, for the selected instances in Fig. 3b, the first threshold contributes more than the second.

In addition, in a GCV plot it is much easier to find patterns and clusters compared to ICE plots. Such expert-guided subgroup discovery can, for instance, be used to assess model fairness, and to discover different ‘strategies’ a model has for predicting the same class. There are two reasons for this.



(a) *ICE plot*: Due to the huge variance in class probabilities it is difficult to find patterns.

(b) *GCV plot*: Selected polyline (blue), revealing two clusters with diverging contribution values.

Fig. 3 Two visualizations of a random forest (100 trees) trained on the Wine Quality dataset (Cortez et al. 2009), showing feature "pH"

First, in ICE plots the differences in prediction probability lead to vertical dispersion of polylines that obscures global patterns. For example, both plots in Fig. 3 reflect the same model and data. The GCV plot in Fig. 3b clearly highlights two different clusters (the selected and non-selected lines), whereas this bi-modality is difficult to spot in Fig. 3a. The lower vertical dispersion in GCV plots also enables intuitive interactive selection by means of lasso brushing (Raidou et al. 2015), as shown in Fig. 3b. In an ICE plot, lasso selection does not yield any interesting clusters; this would require selecting lines based on angle.

To address the vertical dispersion, Goldstein et. al. discuss a variant called centered ICE plots that center the curves at a certain feature value x_S and display only the difference in prediction to this point. However, some dispersion remains, finding a suitable value for x_S is challenging, and the interpretation of the y-axis becomes very unclear. In addition, the authors introduce a derivative ICE variant. This approach is similar to a GCV plot using LIME, but considers only the derivative with respect to a single feature, whereas LIME considers all features.

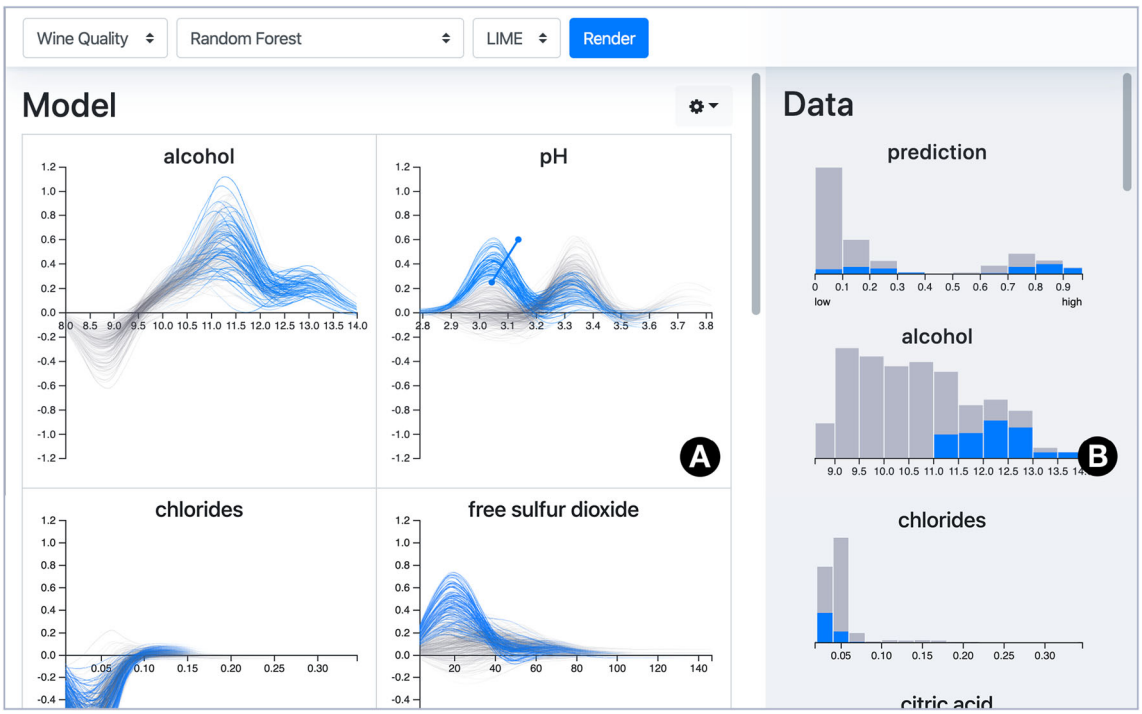


Fig. 4 Overview of our visual analytics instrumentation of techniques in Fig. 1. More features are revealed by scrolling down. Line fading is enabled with $\tau = 0.2$

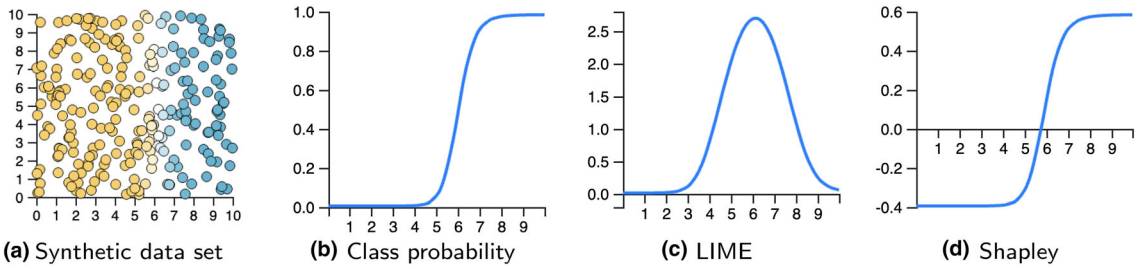


Fig. 5 The class probability **b** of a model trained on a synthetic dataset **a** and two LCV plots using different contribution techniques **c**, **d**. All plots share the same *x*-axis domain

Second, feature contribution techniques have to simplify in order to approximate the reference model. For instance, LIME fits a linear model to a sampling region around an instance. This simplification yields smooth curves in LCV and GCV plots, making it easier to spot more subtle patterns (Graham and Kennedy 2003). This also gives an intuitive visual interpretation of the kernel size parameter in LIME: changing this parameter affects the smoothness of the curves.

5 Design

We built a visual analytics instrumentation of all discussed techniques (Fig. 1), as they are valuable in different situations. More detail and a usage scenario are shown in the supplemental material. It can be used by data scientists to understand how a feature impacts model predictions on a global level. In addition, it also shows what values of a feature are relevant. Through interaction, different patterns in feature contribution can be analyzed.

5.1 Feature contribution technique

Even though feature contribution techniques can provide great insight into model predictions, the output of different techniques may vary significantly, making it challenging to compare them.

The examples in Fig. 5 show that LCV plots with different explanation techniques can vary significantly. The difference is that LIME contribution values are approximate (partial) derivatives, whereas Shapley contributions are additive: the sum of all feature contributions (plus the constant *base rate*, i.e., the average predicted value) equals the class probability $\hat{y}$. To further explain the difference, we consider the relation between contribution vectors and the class probability for the various methods:

$$\begin{aligned} \text{LIME:} \quad & \hat{y} = \alpha + \sum_i \beta_i X_i \\ \text{Shapley:} \quad & \hat{y} = \epsilon + \sum_i \phi_i \end{aligned} \tag{5}$$

Because the base rate ϵ of Shapley values is constant, the sum of Shapley contribution values ϕ recovers the original class probability $\hat{y}$. For LIME, the contribution values need to be composed with the feature values first. Next, the linear regression intercept (α) is not constant but varies per instance.

For this paper, we focus on LIME contribution as it has a more straightforward interpretation (i.e., which small change in feature value results in a big change in prediction) than Shapley values, and a lower computational cost (Garreau and von Luxburg 2020). A kernel size of 0.5 was used, but we encourage tweaking this parameter on a per-dataset basis.

5.2 Visual encoding

In PDP and LCV plots, a single instance is traced over the entire marginal distribution of a feature. This may yield data points that are out-of-distribution (e.g., a person with age 5 and height 200cm). Such data points force the model to extrapolate to an unseen part of the feature space, which could be misleading.

To account for this, we gradually fade out polylines as they get further away from the original data point. Any kernel can be applied, but in our implementation we use a triangular kernel:

$$\alpha(u) = \max\left(0, 1 - \frac{|u|}{\tau R_t}\right) \tag{6}$$

where τ is a configurable parameter impacting the length of the fade and R_t the range of the marginal distribution of feature t.

The result is shown in Fig. 4a, which depicts the same data as in Fig. 3b. Note that toward the end of the feature range, Fig. 3b shows a third bump in feature contribution. This bump is not visible in Fig. 4a with line fading. This shows that the effect was extrapolated from out-of-distribution data. Additionally, the original data points can be shown to further enable the identification of out-of-distribution effects (Fig. 6a).

Our implementation (shown in Fig. 4) contains two views. The *model view* shows small multiples of GCV plots for all features (Fig. 4a). This enables data scientists to determine which features are used by the model, and what values play an important role in predictions. The y-axis is shared across all plots for easy comparison. Line fading can be customized by configuring the fading parameter τ on-the-fly, and an option is provided to average all local polylines (similar to global PDPs). Selection is enabled by lasso brushing (Raidou et al. 2015): dragging a line in the plot will select all polylines which intersect that line, revealing clusters in the feature contribution vectors. This selection is linked to all other GCV plots and the data view.

The *data view* (Fig. 4b) contains a list of histograms to show the original data distributions. The distribution of the selected instances in the model view is highlighted in blue. In the example, the data view shows that the selected cluster in the model view (for which 'pH'= 3.05 is important to the predictions) corresponds with data instances with high alcohol content. The x-axis of the histograms can also be brushed to selected instances with specific feature values and to highlight lines in the GCVs.

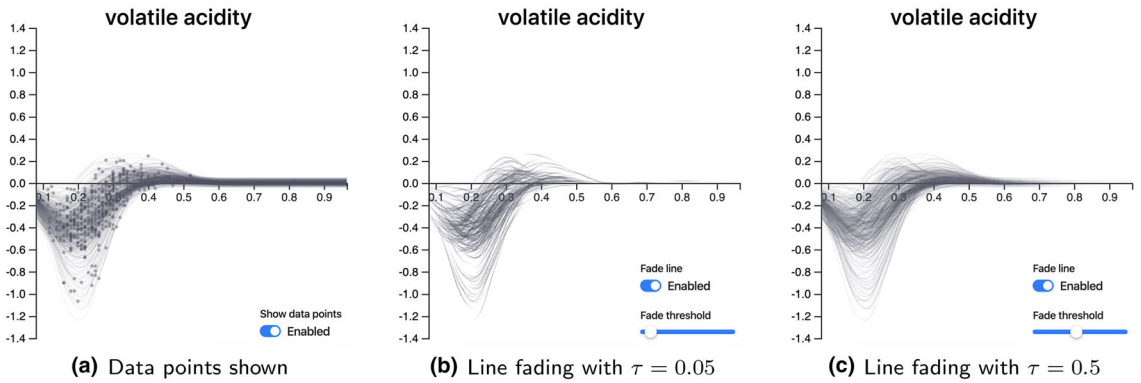
6 User study

To validate our proposed technique, we conduct a comparative evaluation through a user study with a variety of machine learning professionals and visualization experts. The goal of the experiment is to analyze how experts use different visualizations to understand complex machine learning models. We aim to answer the following research questions:

RQ1 Can experts determine which features are most relevant and most used by the model for predictions?
RQ2 Does the visualization enable the understanding of the relationship between feature value and importance? Can experts find feature values at which the prediction changes drastically?
RQ3 Are experts enabled to detect divergent model behavior (i.e., groups of instances are treated differently)? This corresponds to different 'strategies' the model employs to produce predictions.

6.1 Participants

We invited 66 experts with an interest in machine learning explanations. We received 22 replies, of which 6 were female, 15 male and 1 other. The ages of the participants range from 24 to 50 years. Ten participants reported having high experience with machine learning (>3 on 5-point Likert scale), while 9 participants



(a) Data points shown **(b)** Line fading with $\tau = 0.05$ **(c)** Line fading with $\tau = 0.5$

Fig. 6 Included methods to enable the identification of out-of-distribution effects

reported having high experience with visualization (>3 on 5-point Likert scale). Seven participants reported having used explainable AI techniques such as LIME and SHAP before, the rest was new to the concept. Participants were not compensated for their contribution.

6.2 Study procedure

We set up an interactive online survey that took around 10 to 20 minutes to complete. To start, each participant completed a short background survey we used to report on the population demographics. Participants were then introduced to the different visualizations for model interpretability as listed in Fig. 1 and to the Wine Quality dataset. Finally, the participants were asked three sets of 10 questions about a complex model predicting wine quality. Each set corresponds to one of the main research questions and is preceded with an introductory example. To compare the different techniques, the participants were provided with ICE plots for the first five questions of each set and GCV plots for the latter half. To avoid a learning effect due to the order of the plots, the features used were distinct.

RQ1 Participants were presented with visualizations of two randomly selected features of the Wine Quality dataset and had to indicate which of the two was more relevant to the model predictions.
RQ1 Participants were shown a visualization of a single feature and had to indicate the most important feature value (i.e., for which feature value the prediction changed most rapidly).
RQ1 Participants were enabled to use lasso selection and were tasked to detect whether certain wines were treated differently by the model than others (i.e., whether model strategies exist).

6.3 Results

We recorded the answers, the self-reported confidence in the answer on a 5-point Likert scale, and the time spent at each question in milliseconds. The participants successfully completed the questions in 12 minutes on average, excluding the background survey and introductory example.

To test for statistical significance, we will use one-sided proportion Z-test for the proportion of correct vs. incorrect answers, the Mann–Whitney test for self-reported confidence due to the ordinal nature of the Likert scale, and the t-test for the time taken for each question. For the alternative hypotheses, we assert that participants have a higher proportion of correct answers, higher confidence, and less time taken using GCV plots. We evaluated different participation cohorts independently, but did not find a significant difference. Furthermore, the number of participants in each cohort is insufficient to prove statistical significance.

6.3.1 RQ1. Feature importance

In general, participants were able to determine which of the two presented features was most important using both visualizations. We define the correct answer as the average LIME contribution for both features; the one with the largest average contribution was deemed most important. On average, 69.1% of participants selected the correct answer using ICE plots and 86.3% using GCV plots. This result is statistically significant with $p = 0.001$.

In addition, participants were a lot more confident in their answers with GCV plots with a statistical significance of $p = 8.91e\text{-}6$. The distributions of reported confidences for each technique are shown in Fig. 7a.

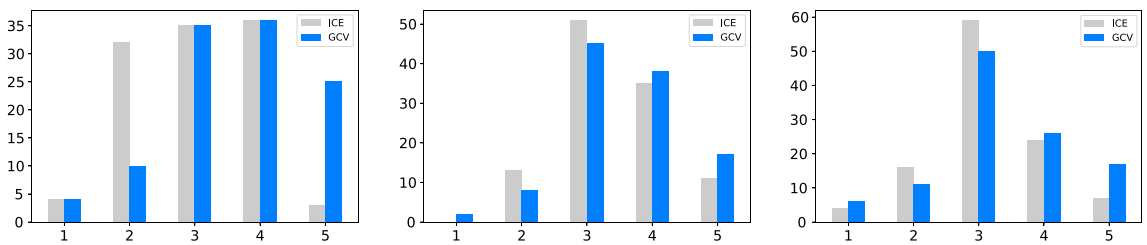


Fig. 7 Self-reported confidence per research question, separated by ICE plot (grey) and GCV plot (blue). Mann–Whitney significance test p-values are annotated

Finally, participants took less time finding an answer: 21.9 s with ICE plots and 10.6 s with GCV plots. As participants took a bit longer on the first question for each visualization, we omitted it from the time averages.

6.3.2 RQ2. Feature value and contribution relationship

Participants also understood the relationship between feature value and importance. We defined the correct answer as the feature value with the highest average LIME contributions out of all line segments. We deemed the answer correct if the participants were within 5% of the correct answer (relative to the marginal range of the feature) to account for insignificant deviations and tick value bias. On average, 56.4% of participants selected the correct feature value using ICE plots and 75.5% using GCV plots. This result is statistically significant with $p = 0.0014$.

We found a slight positive difference between reported confidences. However, this difference does not pass the statistical test ($p = 0.084 > 0.05$). We hypothesize this is, at least in part, due to participants being less rigorous when reporting confidence: for RQ2 and RQ3 a few participants seemed to have just left the slider on the default value. The distributions are shown in Fig. 7b.

In terms of taken time there is again a significant difference of 14.7 s for ICE plots and 11.4 s for GCV plots ($p = 0.0038$). The first questions for each visualization technique were again omitted.

6.3.3 RQ3. Model strategies

Finally, we tested whether participants were able to discern different model strategies. Unfortunately, there was no statistical difference ($p = 0.3314$) in correctness: 67.3% of participants selected the correct answer using ICE plots and 70% using GCV plots. This may be caused by the relatively simple dataset used for the experiment, making it easier to spot strategies in ICE plots regardless of the many occluding and intersecting lines. Another contributing factor is the ambiguity of what constitutes a cluster. The provided example may have been insufficient to explain the concept of model strategies we expected.

In terms of confidence (Fig. 7c) and time taken, there was again a statistically significant difference with p-values of 0.0447 and 0.0013, respectively. Using ICE plots, participants took 27.2 s and with GCV plots only 13.9 s.

6.3.4 Feedback

At the end of the experiment, participants were asked to provide optional comments about the survey. We received three questions about how best to interpret importance in an ICE plot. As we have argued, the translation of variation in the predicted value shown in ICE plots is subjective and challenging, whereas GCV plots directly show feature importance by value, adopting the assumptions of the underlying explanation technique. As a result, all these participants reported higher confidence using GCV plots over ICE plots during RQ1.

Regarding RQ3, one participant remarked that using the interaction helped them find the clusters. Two other participants mentioned they liked this part of the survey, but found it difficult to determine what constitutes a cluster. This is valid feedback and reflected in the lack of significant correctness results for RQ3.

7 Discussion and future work

Our proposed visualization supports answering various questions about the model to understand a complex model. First, an expert can check the feature contribution and relationship with feature value at a single glance. In prior work, this could only be done with separate visualizations of feature contribution and partial dependence-based plots. We showed these are difficult to compare (it requires estimating the slope), and may not show consistent results, as they encode different information. Next, patterns (or ‘strategies’) can be spotted that would otherwise remain hidden (e.g., Fig. 3b highlights two distinct clusters of lines). In addition, linking with the data view helps to ascertain what constitutes this strategy (e.g., alcohol contents). Finally, our approach enables the validation of the (un)certainty of contribution through line fading.

Our user study has shown that GCV plots can aid the understanding of complex models by increasing correctness and confidence and reducing the time taken to obtain an insight into how complex machine learning models work, compared to traditional techniques.

However, the current implementation has a few limitations. First, much computation is needed to obtain these curves: our examples with all features of the Wine Quality dataset took 5 minutes (on AMD Ryzen 5 3600X); it will take longer for larger datasets and more complex models. Hence, computing these plots on-the-fly is not possible. We address this by caching the results in our implementation. The optimization of current implementations of feature contribution methods for large datasets is an interesting direction for future research.

Next, even though we can visually represent many features there is a practical limitation on the number of features that can be shown. Hence, the plots are best applicable to datasets with at most 10-20 features. In addition, for the plots to be interpretable we rely on a dataset that has features with inherent meaning.

Finally, our work relies on the validity of the used underlying explanation technique. This is not rock solid yet, as both LIME and Shapley values have been criticized (Garreau and von Luxburg 2020; Kumar et al. 2020; Merrick and Taly 2019). We chose LIME as it has a more straightforward interpretation than Shapley values (i.e., which small changes in feature value result in a big change in prediction) and is computed faster. However, we think our plots are able to help experts understand the differences between explanation techniques, ultimately encouraging this line of research.

As a follow-up, the user study can be expanded to cover a wider variety of datasets and participant cohorts to further investigate the suitability of our approach.

8 Conclusion

We have presented local contribution-value (LCV) plots, a novel way of conveying feature contribution as a function of feature values. This was previously only possible by combining multiple views, or by fallibly estimating the slope of partial dependence curves, which is challenging and subject to errors. Furthermore, we introduced global contribution-value (GCV) plots to show a comprehensive overview of the full model behavior. These plots are information dense and enable novel insights into a model. We have addressed uncertainty of the sensitivity analysis by interactively fading out lines, enabling the validation of patterns for real data, and empower an analysis workflow with linked views.

In a user study with 22 machine learning professionals and visualization experts, we have shown that the visualizations support model interpretation by increasing correctness and confidence and reducing the time taken to obtain an insight compared to previous techniques.

The proposed visualizations provide data scientists with an in-depth view of the role of a feature in predictions and enable model diagnosis, refinement, decision support and justification use cases commonly driven by model interpretability (Collaris and van Wijk 2020a).

Acknowledgements 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

Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Müller KR (2010) How to explain individual classification decisions. J Mach Lear Res 11(Jun):1803–1831

Cleveland WS, McGill R (1985) Graphical perception and graphical methods for analyzing scientific data. Science 229(4716):828–833

Collaris D, van Wijk JJ (2020a) ExplainExplore: Visual exploration of machine learning explanations. In: 2020 IEEE Pacific Visualization Symposium (PacificVis), IEEE

Collaris D, van Wijk JJ (2020b) Machine learning interpretability through Contribution-Value Plots. In: Proceedings of the 13th International Symposium on Visual Information Communication and Interaction (VINCI 2020), pp 1–5
Cortez P, Cerdeira A, Almeida F, Matos T, Reis J (2009) Modeling wine preferences by data mining from physicochemical properties. Decis Support Syst 47(4):547–553
Doshi-Velez B Finale; Kim (2017) Towards a rigorous science of interpretable machine learning. In: arXiv:1702.08608
Friedman JH (2001) Greedy function approximation: A gradient boosting machine. Annals of Statistics 29(5):1189–1232, https://doi.org/10.1214/aos/1013203451, arXiv:1011.1669v3
Garreau D, von Luxburg U (2020) Explaining the explainer: A first theoretical analysis of lime. arXiv preprint arXiv:200103447
Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 24(1):44–65
Graham M, Kennedy J (2003) Using curves to enhance parallel coordinate visualisations. In: Proceedings of the 7th international conference on information visualization, IV 2003., IEEE (2003), pp 10–16
Guidotti R, Monreale A, Turini F, Pedreschi D, Giannotti F (2018) A survey of methods for explaining black box models. arXiv preprint arXiv:180201933
Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4401–4410
Kononenko I et al (2010) An efficient explanation of individual classifications using game theory. J Mach Learn Res 11(Jan):1–18
Krause J, Perer A, Bertini E (2016a) Using visual analytics to interpret predictive machine learning models. ICML Workshop on Human Interpretability in Machine Learning pp 106–110, arXiv:1606.05685v1
Krause J, Perer A, Ng K (2016b) Interacting with predictions: Visual inspection of black-box machine learning models. ACM Conf on Human Factors in Computing Systems pp 5686–5697, https://doi.org/10.1145/2858036.2858529
Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler S (2020) Problems with shapley-value-based explanations as feature importance measures. arXiv preprint arXiv:200211097
Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in neural information processing systems, pp 4768–4777
McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GC, Darzi A et al (2020) International evaluation of an ai system for breast cancer screening. Nature 577(7788):89–94
Merrick L, Taly A (2019) The explanation game: Explaining machine learning models with cooperative game theory. arXiv preprint arXiv:190908128
Raidou RG, Eisemann M, Breeuwer M, Eisemann E, Vilanova A (2015) Orientation-enhanced parallel coordinate plots. IEEE Trans Vis Comput Graph 22(1):589–598
Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 1135–1144
Štrumbelj E, Kononenko I, Šikonja MR (2009) Explaining instance classifications with interactions of subsets of feature values. Data Knowl Eng 68(10):886–904
Wexler J, Pushkarna M, Bolukbasi T, Wattenberg M, Viégas F, Wilson J (2019) The what-if tool: interactive probing of machine learning models. IEEE Trans Vis Comput Graph 26(1):56–65
Zhao X, Wu Y, Lee DL, Cui W (2018) iForest: interpreting random forests via visual analytics. IEEE Trans Vis Comput Graph 25(1):407–416