Considerations when learning additive explanations for black-box models

Sarah Tan¹ · Giles Hooker² · Paul Koch³ · Albert Gordo¹ · Rich Caruana³

Received: 19 October 2021 / Revised: 31 December 2022 / Accepted: 30 March 2023 / Published online: 19 June 2023 © The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2023

Abstract
Many methods to explain black-box models, whether local or global, are additive. In this paper, we study global additive explanations for non-additive models, focusing on four explanation methods: partial dependence, Shapley explanations adapted to a global setting, distilled additive explanations, and gradient-based explanations. We show that different explanation methods characterize non-additive components in a black-box model’s prediction function in different ways. We use the concepts of main and total effects to anchor additive explanations, and quantitatively evaluate additive and non-additive explanations. Even though distilled explanations are generally the most accurate additive explanations, non-additive explanations such as tree explanations that explicitly model non-additive components tend to be even more accurate. Despite this, our user study showed that machine learning practitioners were better able to leverage additive explanations for various tasks. These considerations should be taken into account when considering which explanation to trust and use to explain black-box models.

Keywords  Black-box models · Additive explanations · Model distillation · Interaction effects · Correlated features

1 Introduction
Whether explicitly or implicitly, many model explanation methods, whether local or global, require characterizing the impact of a particular feature on the machine learning model’s prediction function. Many of these explanations are additive, explaining model predictions...
by adding up the impact of different features. For example, Shapley additive explanations (Lundberg & Lee, 2017), a local feature importance method, consists of Shapley values that add up to the prediction for a particular point. LIME (Ribeiro et al., 2016) fits a locally linear — hence locally additive — model in the neighborhood around the point, with the weights from the linear model describing how features impact the prediction.

However, when a model’s prediction function has non-additive components, there is no one unique additive explanation that can perfectly characterize it (Hastie & Tibshirani, 1986). Even when models only have additive components, but different components are correlated (Amodio et al., 2014), there is no one unique additive explanation. This has been demonstrated in several practical settings; Orlenko and Moore (Orlenko & Moore, 2021) simulated genetic datasets with non-additive epistatic interactions of different complexity, finding that while feature importance measures, including SHAP, were able to identify genes with the largest independent main effects, different feature importance measures did not judge genes involved in non-additive interactions similarly. Lengerich et al. (2020) found on the MIMIC-III dataset that the feature-prediction relationship for blood urea nitrogen (BUN), a feature that participated in a large number of discovered interactions, was significantly altered after performing a procedure proposed in their paper to move main effects out of interaction terms.

Hence, we are left with a conundrum — when there exists multiple additive explanations that look different from each other, how do we know which additive explanation method to trust and use?

We study this question from several angles in this paper. First, we show how several popular additive explanation methods — partial dependence (Friedman, 2001), Shapley additive explanations (Lundberg & Lee, 2017), distilled additive explanations (Tan, 2018), and gradient-based additive explanations — differentially characterize main effects of features in the presence of non-additive interaction effects that involve these features, or correlation between features. We use the concepts of main and total effects (related to marginal plots (Hastie et al., 2009)), to anchor these additive explanations, and show how each of them distributes interaction effects in a different way. Secondly, we quantitatively compare different additive explanations on several regression and classification tasks, finding that distilled explanations are generally the most accurate additive explanations (unsurprisingly, since their training optimizes for faithfulness to the black-box model) and that non-additive explanations that explicitly model interactions tend to be the most accurate over all.

Despite these considerations, our investigations in this paper find that non-additive explanations are not a panacea. In a user study we performed to evaluate the interpretability of additive explanations compared to non-additive explanations that may not suffer from these same considerations, we found that machine learning practitioners were able to leverage additive explanations for various tasks, neither trusting the explanation less due to perceived oversimplification (as in the case of linear explanations) or being overwhelmed by the complexity of the explanation (as in the case of large decision trees).

The contributions of this paper are:

- Proposing distilled additive explanations. While distilled tree-based explanations have been proposed in the literature (Craven & Shavlik, 1995), additive models had not been tried as distilled students, and additive explanations were typically learned using permutation-based approaches such as partial dependence or Shapley values. Our paper presents insights uncovered when learning additive explanations using distillation.
- A comprehensive study of leading additive explanation methods, writing them in the same mathematical framework so they can be compared to each other, investigating how they perform under key settings that not all methods, when first proposed, considered (non-
additive interactions between features, correlation between features). By studying these explanations not just empirically on real data, but also comparing them to the theoretically grounded main and total effects that can be calculated from synthetic functions, our analysis provides a fresh view into the differences between these explanation methods beyond simply fidelity and accuracy of an explanation.

- To the best of our knowledge there has not been any user studies with expert users (like ours) that compare different classes of global explanations, in particular global additive and global non-additive explanations; our paper is the first to compare the interpretability of explanations that are visually very different.

The rest of the paper is organized as follows. Section 2 discusses related work. In Sect. 3 we review several popular explanation methods and show how they are additive. We consider how additive explanations are impacted by interaction effects and correlation in Sect. 4, and provide results of empirical investigations of these phenomena in Sect. 5. In Sect. 6 we describe results from the user study. Section 7 concludes the paper.

2 Related work

2.1 Global explanations

Post-hoc explanations of black-box models can be roughly divided into local methods that explain predictions at individual points and global methods that explain overall model predictions (Doshi-Velez & Kim, 2018). Examples of global explanations include using a second interpretable model to approximate the black-box model, often termed “model distillation” (Fu, 1994; Craven & Shavlik, 1995; Sanchez et al., 2015; Frosted & Hinten, 2018; Bastani et al., 2017; Lakkaraju et al., 2019; Ibrahim et al., 2019), analyzing intermediate representations (Bau et al., 2017; Mu & Andreas, 2020) or concepts (Kim et al., 2018) encoded by the black-box model, prototype selection (Bien & Tibshirani, 2011; Kim et al., 2016; Tan et al., 2020), counterfactual explanations (Rawal & Lakkaraju, 2020) that summarize actions that can be taken to change the black-box model’s predictions for an entire population, and feature importance measures (Breiman, 2001; Lundberg & Lee, 2017; Fisher et al., 2019; Covert et al., 2020; Williamson & Feng, 2020; Yan & Procaccia, 2021). We focus on global post-hoc explanations in this paper, and investigate how many of these explanation methods rely on approximating non-additive functions additively.

2.2 Trustworthiness of explanations

One viewpoint in the community is that post-hoc explanations of black-box models can be unreliable or misleading (Rudin, 2019), and that inherently interpretable models should be favored. For example, Slack et al. showed that one can carefully craft black-box models that lead to innocuous explanations while still producing biased outputs (Slack et al., 2020). A known problem of additive explanations that affects their trustworthiness is their multiplicity. Main effects and interaction effects can freely move around while producing the same prediction (Lengerich et al., 2020). In addition, different training strategies may lead to significantly different or even contradictory models (Chang et al., 2021). In this paper we empirically investigate this phenomenon on additive explanations.
2.3 Evaluation of interpretability

There is no universal definition of interpretability (Doshi-Velez & Kim, 2018); many recent papers evaluate interpretability in terms of how a human uses the model to perform downstream tasks. These studies are typically performed on non-expert humans (e.g., Mechanical Turkers) (Lage et al., 2019; Poursabzi-Sangdeh et al., 2021); recently, there has been more work evaluating interpretability on expert users (Lakkaraju et al., 2019; Kaur et al., 2020; Jesus et al., 2021). A contribution of this paper is a user study comparing global additive and non-additive explanations on expert users.

3 Background

In this section, we review definitions and concepts that will be used throughout this paper.

3.1 Global and local additive explanations

Let $\mathbf{X}$ be a $n \times p$ matrix consisting of feature values of points $\{x^j_i, i = 1, \ldots, p, j = 1, \ldots, n\}$ where $p$ is the number of features and $n$ is the number of points. Let the vector $X^j = (x^j_1, \ldots, x^j_p)$ represent all feature values of the $j$th point, and let the vector $X_i = (x^1_i, \ldots, x^n_i)$ represent the $i$th feature value across all points.

Definition 3.1 (Global additive explanations) Given $F$, the prediction function of a black-box model, global additive explanations $G$ decompose $F$, as a sum of effects of $p$ features:

$$G(\mathbf{X}) = h_0 + \sum_{i=1}^{p} h_i(X_i) + \sum_{i=1}^{p} \sum_{i'=1}^{p} h_{ii'}(X_i, X_{i'}) + \ldots$$ (1)

and we measure the accuracy of the representation $G(\mathbf{X})$ by $||F(\mathbf{X}) - G(\mathbf{X})|| = \text{error}_G$. $h_0$ is a bias term, $h_i(X_i)$ is the “main effect” of feature $X_i$ on $G$ while $h_{ii'}(X_i, X_{i'})$ is a second-order “interaction effect” of features $X_i$ and $X_{i'}$ on $F$ (second-order as it consists of two features). This definition of $G$ also accommodates higher-order interaction effects, i.e. third-order and beyond. Such a decomposition is not unique and each of the methods examined below will give different results. We will refer to main effects and interactions in terms of algebraic groupings rather than how those groupings were derived.

Definition 3.2 (Local additive explanations) Given $F^{(X^j)}$, the prediction function of a black-box model on a single point $X^j = (x^j_1, \ldots, x^j_p)$, local additive explanations $\Gamma^j$ decompose $F(X^j)$, as a sum of effects of $p$ features:

$$\Gamma^j(X^j) = \phi_0^j + \sum_{i=1}^{p} \phi_i^j(x^j_i)$$ (2)

In Sect. 3.2.2 we show how global additive explanations can be constructed from multiple local additive explanations.

Visualization and interpretability. Global additive explanations can be visualized as in Fig. 1 where we produce one plot per feature, where each plot’s x-axis is the domain of input feature $X_i$ and the y-axis is the feature’s contribution to the prediction $h_i(X_i)$. Partial dependence
plots (Friedman, 2001) and Shapley dependence plots (Lundberg & Lee, 2017) are visualized this way. As the number of features $p$ increases, there are more plots for a human to review and understand, which could decrease the interpretability of such explanations.

### 3.2 Additive explanations

We now review several popular explanation methods and show how they can be written as additive equations.

#### 3.2.1 Partial dependence and marginal plots

The partial dependence (Friedman, 2001) for feature $X_i$ is defined as:

$$q_i(X_i) = E_{X_C}[F(X_i, X_C)] = \int F(X_i, X_C)dP(X_C)$$

where $X_C$ is the set of all features excluding the $i$th feature. In practice, $q_i(X_i)$ is the average prediction from the data set when the value $X^j_i$ is replaced with the $X_i$ of interest. Specifically, for a specific value $z$:

$$\hat{q}_i(X_i = z) = \frac{1}{n} \sum_{j=1}^{n} F(x_1^j, \ldots, x_{i-1}^j, x_i^j = z, x_{i+1}^j, \ldots, x_p^j)$$

Partial dependence plots of different features are often interpreted as being summed up as in Eq. 1 with $h_i(X_i) = q_i(X_i)$ to construct a global additive explanation. However, to make this precise, the $q_i(X_i)$ need to be centered. This is because each $q_i(X_i)$ averages the values of $F$ and their sum will average to approximately $p$ times the average of $F$. To center them, we subtract $\overline{q}_i = \frac{1}{|z|} \sum_z \hat{q}_i(X_i = z)$, the average of $q_i(X_i)$ over all possible values of $z$, to obtain the global additive explanation:

$$PD(X) = \hat{h}_0 + \sum_{i=1}^{p} {\hat{q}_i(X_i) - \overline{q}_i}$$

where $\hat{h}_0 = \frac{1}{n} \sum_j F(X^j)$. Partial dependence takes the expectation of the prediction function over the marginal distribution of $X_C$, independent of the value of $X_i = z$. Alternatively, this expectation could also be taken over the conditional distribution $X_C | X_i = z$. This is the marginal plot1 (Hastie et al., 2009; Zhao & Hastie, 2021; Apley & Zhu, 2020), defined as:

$$r(X_i) = E_{X_C|X_i}[F(X_i, X_C) | X_i] = \int F(X_i, X_C) P(X_C | X_i) dX_C$$

In practice, marginal plots are constructed by finding the collection of points for which $X_i = z$, $N(z) = \{j : X^j_i = z\}$ and averaging over these for each $z$:

$$\hat{r}_i(X_i = z) = \frac{1}{|N(z)|} \sum_{j \in N(z)} F(X^j)$$

1 The name is counter-intuitive as it is based on the conditional, not marginal distribution.
Fig. 1 From local Shapley explanations to gSHAP, a global Shapley additive explanation for a particular feature created by aggregating local Shapley explanations.

Marginal plots of different features can also be centered and summed up to construct a global additive explanation:

$$M(X) = \hat{h}_0 + \sum_{i=1}^{p} \{\hat{r}_i(X_i) - \hat{r}_i\}$$ (4)

where $\hat{r}_i$ is constructed using a similar centering procedure on $r_j(X_j)$ as was performed on partial dependence plots. Note that while partial dependence plots will recover components of an underlying additive function, marginal plots will only do so if the features are all independent.

### 3.2.2 Shapley additive explanations

The Shapley local additive explanation (Lundberg & Lee, 2017) for a point $X^j$ consists of $\phi^j_i$ that are defined to sum to $F(X^j)$ up as in Eq. 2. To achieve this, we need to define prediction functions $F_S(X_S)$ for each subset $S$ of the features. Lundberg and Lee (2017) defines $F_S(X_S) = \int F(X_S, X_C)P(X_C)dX_C$ as in partial dependence plots, but alternative formulations are also possible. The Shapley value for feature $X_i$ at point $X^j$ is then given by:

$$\phi^j_i(x^j_i) = \sum_{S \subseteq \{x^j_1, ..., x^j_p\} \setminus x^j_i} \frac{|S|!(p - |S| - 1)!}{p!} \left[ F_{S \cup \{i\}}(X^j_{S \cup \{i\}}) - F_S(X^j_S) \right]$$ (5)

For example, each blue point in Fig. 1 is one such $\phi^j_i$ for feature $i =$ “Temperature” and a particular point $j$. In practice, Eq. 5 is computed by approximation, using techniques that repeatedly query the black-box model such as Kernel SHAP (Lundberg & Lee, 2017), Shapley sampling values ((Štrumbelj & Kononenko, 2014).

Note, importantly, that while $F(X) = \sum \phi^j_i(X^j_i)$ at each point $X^j$, this does not directly define an additive model in the sense of (1) since $\phi^j_i(X^j_i)$ changes with features other than $X_i$. Here we aggregate local explanations that act on single points to construct a global explanation.
Bhatt et al. (2020); Setzu et al. (2021); van der Linden et al. (2019). For example, the red line in Fig. 1 was created using the basic approach of feature-wise sample mean (Bhatt et al., 2020), averaging $\phi_j^{\text{Temperature}}$ at each unique temperature value $z$ to construct a global additive explanation that we call gSHAP and define as:

$$gSHAP(X) = \hat{h}_0 + \frac{1}{n} \sum_{j=1}^{n} \phi_i^j \left(x_i^j\right)$$ (6)

While gSHAP is not the only way to construct a global explanation from Shapley local explanations, it is intuitive and easy to implement on top of Shapley local additive explanations, hence we experiment with it in this paper. Another example of global Shapley values is SAGE (Covert et al., 2020) which decomposes a model’s accuracy function, but not prediction function, into Shapley values of each feature.

### 3.2.3 Distilled additive explanations

Different from the previous methods that construct explanations by repeatedly querying the black-box model, this method treats the black-box model as a “teacher” model and uses model distillation techniques (Bucilua et al., 2006; Ba & Caruana, 2014; Hinton et al., 2014) to train a white-box model (“student” model) that approximates and explains $F$. Different white-box model classes can be as the student model, such as decision trees (Craven & Shavlik, 1995; Frost & Hinton, 2018), generalized additive models (Tan, 2018), linear models, etc. If an additive model is used, the resulting explanation can be written as:

$$AM(X) = \hat{h}_0 + \sum_{i=1}^{p} \hat{h}_i(X_i)$$ (7)

where the $\hat{h}_i$ are obtained by minimizing the mean-squared error between the teacher model’s prediction function $F$ and student model’s explanation $AM$ on training samples:

$$\hat{h}_0, \hat{h}_1, \ldots, \hat{h}_p = \arg\min L(h_0, h_1, \ldots, h_p) = \arg\min \frac{1}{n} \sum_{j=1}^{n} \| F(X^j) - G(X^j) \|_2^2$$

$$= \arg\min \frac{1}{n} \sum_{j=1}^{n} \left| F(X^j) - \left( \hat{h}_0 + \sum_{i=1}^{p} \hat{h}_i(x_i^j) \right) \right|_2^2$$

In this paper we consider two types of additive models: spline-based (Wood, 2006) and tree-based (Lou et al., 2012) whose resulting explanations we call Student Additive Splines (SAS) and Student Bagged Additive Boosted Trees (SAT). For SAS, $\hat{h}_i(X_i)$ is a cubic regression spline function of feature $X_i$ implemented using the mgcv library (Wood, 2011). For SAT, $\hat{h}_i(X_i)$ is a bagged tree function of feature $X_i$ implemented using the InterpretML (Nori et al., 2019) package. Pairwise interactions can also be modeled explicitly (Lou et al., 2013), in the form of $h_{ii'}(x_i^j, x_i'^j)$ added to Eq. 7. We call this resulting explanation SAT + pairs.
3.2.4 Gradient-based explanations

Finally we consider a gradient-based explanation that involves constructing an additive function $g_{GRAD}$ through the first-order Taylor decomposition of $F$:

$$g_{GRAD}(X) = F(0) + \sum_{i=1}^{p} \frac{1}{n} \sum_{j=1}^{n} \frac{\partial F(X_j)}{\partial x_i} x_i^j$$  \hspace{1cm} (8)

This is related to the “gradient*input” method (e.g. (Shrikumar et al., 2017)) used to generate saliency maps for images. While there are several other gradient-based feature attribution methods ((Simonyan et al., 2014; Simonyan & Zisserman, 2015; Bach et al., 2015; Shrikumar et al., 2017); also see (Montavon et al., 2018) or Ancona et al. (2018) for a review), the “gradient*input” method is simple, intuitive, and studied a lot in the literature (Ancona et al., 2018; Adebayo et al., 2019), hence we experiment with it.

4 Non-additive interaction effects and correlation

We now consider how additive explanations are impacted by non-additive interaction effects and correlation between features. For ease of exposition, we will assume that the black-box model’s prediction function $F$ can be written as:

$$F(X) = e_0 + \sum_{i=1}^{p} e_i(X_i) + \sum_{i=1}^{p} \sum_{i' = 1; i' \neq i}^{p} e_{ii'}(X_i, X_{i'}) + \text{error}$$

by decomposing $F$ into successive additive components of increasing complexity and using the error term to capture any remaining effects. Extending to this 2nd-order model allows us to explore the effect of non-additivity on additive explanations.

**Main effect.** A simple way to report the impact of feature $X_i$ on prediction function $F$ is to simply ignore (i) any interaction effects involving $X_i$; (ii) any correlation between $X_i$ and other features. This results in a measure of the impact of feature $X_i$ on $F$ that use only the main effect of $X_i$:

$$M(X_i) = \text{func}(e_i(X_i)) \quad M(X_{i'}) = \text{func}[e_{i'}(X_{i'})]$$  \hspace{1cm} (9)

For example, when $\text{func}$ is the variance function, this is the main effect Sobol index (Sobol’, 1990; Owen, 2014). While main effect measures undercounts the impact of $X_i$ on $F$ in case of (i) this sidesteps the challenging issue of determining how to allocate the interaction effect $e_{ii'}(X_{ii'})$ between features $X_i$ and $X_{i'}$.

**Total effect.** Another simple way to report the impact of feature $X_i$ on $F$ is to attribute an interaction effect to each feature that was part of that interaction effect. This results in a measure of the impact of feature $X_i$ on $F$ that use the main effect of $X_i$ and all interaction effects containing $X_i$:

$$T(X_i) = \text{func}[e_i(X_i)] + \sum_{i' = 1; i' \neq i}^{p} e_{ii'}(X_i, X_{i'})$$

$$T(X_{i'}) = \text{func}[e_{i'}(X_{i'})] + \sum_{i' = 1; i' \neq i}^{p} e_{ii'}(X_i, X_{i'})$$  \hspace{1cm} (10)
For example, when $\text{func}$ is the variance function, this is the total effect Sobol index (Sobol’, 1990; Owen, 2014), and when $\text{func}$ is the expectation function, depending on the domain that the expectation is taken over, this is either the marginal effect (the best least squares approximation to $F$ using only feature $X_i$ (Hastie et al., 2009), also called marginal plots in Apley and Zhu (2020)) or partial dependence plots (Friedman, 2001)). In cases where $\text{func}$ satisfies countable additivity (e.g. the expectation function, where the expectation of the sum equals the sum of the expectations), marginal plots can also be produced for individual interactions or for groups of interactions. While total effect measures overcount the impact of $X_i$ on $F$ in both cases (i) and (ii), this also sidesteps the challenging issue of determining how to allocate the interaction effect $e_{ii'}(X_{ii'})$ between features $X_i$ and $X_{i'}$, albeit in a different way, by allocating it to all features that make up that interaction.

**Distributed effect.** It is clear that neither main nor total effects are the “right” way to allocate an interaction effect between features, though they can still be useful to anchor other additive explanations. In between them are approaches that split up an interaction effect between features, though they can still be useful to anchor other additive explanations. For example, distillation approaches that train additive student models to mimic the prediction functions of non-additive black-box models (Tan et al., 2018) express different features – For example, distillation approaches that train additive student models to mimic the prediction functions of non-additive black-box models (Tan et al., 2018) express non-additive components as a best-fit additive approximation added to additive components. Shapley additive explanations (Lundberg & Lee, 2017) distribute an interaction effect equally among features that made up that interaction. The aim of this paper is to study these explanations in the presence of non-additive interaction effects and correlation between features.

### 4.1 Example

To make this more concrete, we consider an example function for which we have its analytical form and can hence compute the main and total effects for each feature:

$$F_2(x) = F_1(x) + x_1x_2 + |x_3|^2|x_4| + \sec(x_3x_5x_6)$$

where

$$F_1(x) = 3x_1 + x_2^3 - \pi x_3 + \exp(-2x_4^2) + \frac{1}{2 + |x_5|} + x_6 \log(|x_6|) + \sqrt{2 |x_7|} + \max(0, x_7) + x_8^4 + 2 \cos(\pi x_8)$$

This non-additive, highly nonlinear function consists of components from synthetic functions proposed by Friedman and Popescu (2008), Hooker (2004) and Tsang et al. (2018). Like (Tsang et al., 2018), we set the domain of all features to be independent and Uniform[-1,1]. Like (Friedman & Popescu, 2008), we add noise features ($x_9$ and $x_{10}$) to our samples that have no effect on $F_1(x)$.

The **main effect** for feature $x_4$ is $\exp(-2x_4^2)$. The **total effect** (specifically, the marginal plot) of feature $x_4$ is $E[\exp(-2x_4^2) + |x_3|^2|x_4|, x_4] = \exp(-2x_4^2) + \int_{-1}^{1} |x_3|^2|x_4| dP(x_3|x_4) = \exp(-2x_4^2) + \frac{1}{2|x_4|+1}$. This contribution of the interaction term to the impact of feature $x_4$, $\frac{1}{2|x_4|+1}$, manifests as the upward pointing cusp (gray line) in Fig. 3, pointier and higher than main effect alone (black line).

However, for features $x_1$ and $x_2$, their **total effects** are actually the same as their **main effects**, as $E[x_1x_2|x_2] = E[x_1x_2|x_1] = \int_{-1}^{1} x_1 x_2 dP(x_2|x_1) = \frac{x_1 x_2^3}{2}|x_2|^{-1} = 0$, despite the presence of an interaction term $x_1x_2$ in $F_2$. As we will see in Fig. 3, in this case, all additive explanations were not visually different from the main effects only line (black line). Hence,
in this paper, whenever the black-box model function is known, we will provide the main and total effects estimates to ground the additive explanations studied.

5 Experiments

We now conduct empirical investigations of how non-additive interaction effects and correlation between features may impact additive explanations.

**Black box models:** We use multi-layer perceptrons (MLP) neural networks as our black box models, though other models such as gradient boosted trees or random forests could also be used. These models contain non-additive interaction effects, even for very shallow networks, which makes them good candidates for our problem. As is standard, we use ReLU non-linearities between the different layers of the model. We also experimented with using batch normalization (Ioffe & Szegedy, 2015) between layers, but we did not observe consistent gains to justify their use.

Models were trained using cross-entropy for classification problems, and root mean square error (RMSE) for regression ones. We divided the data in train, validation and test splits, where 70% of the data was used for training, 15% for validation, and 15% for test, and used 5 train-validation folds.

We used random search to find the optimal hyperparameters (number of hidden units per layer, learning rate, weight decay, dropout probability, batch size, etc) based on average validation performance on the train-validation splits and multiple random Xavier initializations (Glorot & Bengio, 2010). Following common practice we select powers of 2 for the number of hidden units and batch size, and we search using a logarithmic scale to find the optimal learning rate and weight decay. We use the Adam optimizer (Kingma & Ba, 2015) with default beta parameters, and early stopping based on validation loss. Training was performed in PyTorch.

We experimented with different depths (number of layers), and found that models of a given depth behaved in a consistent manner. In most cases, models with one hidden layer underfit compared to models with two layers, i.e., both their validation and training loss was higher. Models with three or four hidden layers systematically obtained almost-perfect training loss, but their performance on the validation and test sets was not as good, even when strong regularization through dropout and weight decay was applied.

Based on those results, we selected two black box models for our experiments throughout the paper: a 2-hidden layers model with 512 hidden units per layer (2H-512,512), which will be our main model, and a 1-hidden layer model with 8 units (1H-8), which will be our low-capacity model used to evaluate trade-offs between accuracy and interpretability.

**Quantitative evaluation of explanations:** We will quantitatively evaluate explanations as if they were models, following Lundberg and Lee (2017) who suggested viewing an explanation of a model’s prediction as a model itself. Specifically, we evaluate accuracy (how well the explanation predicts the original label). A similar evaluation was performed by Kim et al. (2016) who used their explanations (prototypes) to classify test data.

To learn our explanations we follow a similar approach than to train our black box models: we perform 5 different train-validation data splits (70% of the data for train, 15% for validation and 15% for test) and tune all necessary parameters aiming at maximizing the accuracy on the validation set.
Table 1 RMSE of 1H and 2H models, SAT and PD explanations on $F_1$. The accuracy of a GAM model trained directly on the data is 0.02 RMSE

| Model            | SAT explanation (distilled) | PD explanation |
|------------------|-----------------------------|----------------|
| 1H-8 model       | 0.46                        | 0.46           |
| 2H-512,512 model | 0.14                        | 0.12           |

5.1 Preliminaries: do explanation explain the black-box model or the original data?

Hypothesis: Explanations explain the black-box model behavior, not the original data. Setup: Given a model trained on data produced by a known formula (e.g. $F_1$), we can compare an explanation of the model with the model trained directly on the data. When the explanation of the data is accurate, we can attribute the differences between the explanations to the black-box model behavior, and not to lack of fidelity of the explanations. We use function $F_1$ introduced in Sect. 4, that does not include any interactions between features. Although this function is easily modeled by additive models, neural nets struggle to model it accurately due to how they combine all features through each layer, forcing the introduction of combinations that do not exist in the original function. We simulate 50,000 samples, and train two neural nets models we aim to explain, a 2H-512,512 and 1H-8, to predict $F_1$ from the ten features. The high capacity 2H-512,512 neural net obtained a test accuracy RMSE of 0.14, while the low-capacity 1H-8 neural net obtained a test accuracy RMSE of 0.46, more than 3x worse (cf. Table 1). Next, we train: (1) An additive model trained directly to predict the labels in the data. Since this model is very accurate (RMSE of 0.02), it will help us anchor the explanations. (2) Explanations based on the same additive models, trained to mimic the output from the black-box neural nets. For this experiment, we focus on tree-based additive explanations, but the conclusions also apply to the other additive explanations. (3) Explanations based on partial dependence (similar behavior is observed with Shapley additive explanations).

Results: Fig. 2 focuses on two features of $F_1$ ($x_4$ and $x_6$) and displays: (1) A plot of their analytical expression (in gray). (2) Main effects of an additive model trained directly on the data (in green). (3) Distilled additive explanations of the two neural nets (in blue and cyan). (4) Partial dependence plots for the two neural nets (in red and orange). Their accuracy is shown in Table 1.

We make the following observations:

- Additive models trained to directly predict labels in the data (in green) match almost-perfectly the analytical expression (in gray), showing that additive models have enough capacity to model additive functions without interactions. This models achieves a 0.02 RMSE, a very good value compared to the RMSE achieved by the neural nets (0.14 and 0.46).
- The distilled additive explanation of the 2H-512,512 neural net (in blue) largely matches the analytical expression (in gray), but fails to match some complex parts of the shape, unlike additive models trained to directly predict labels in the data (in green). Quantitatively, this model achieves an RMSE of 0.12, comparable – in fact, better, consistent with other observations regarding model distillation – to the 2H-512,512 teacher, but worse than the 0.02 RMSE achieved by the additive model trained directly to predict the data labels. Since the additive model has enough capacity to almost-perfectly model the data, we conclude the differences observed in the explanation of the 2H-512,512 neural net must be attributed to the neural net behavior.
For two features of function $F_1$: Main effect, GAM trained on data, SAT explanation of complex black-box model (2H-512,512 neural net), PD explanation of the same 2H model, SAT of simpler black-box model (1H-8 neural net), and PD explanation of the same 1H model (Color figure online)

- The distilled additive explanation of the 1H-8 neural net (in cyan) looks notably different than the distilled additive explanation of the 2H-512,512 neural net (in blue), and are also less similar to the analytical expression (in gray). This is consistent with the RMSE error of the 1H model (0.46 RMSE) and the error of the student (also 0.46 RMSE).
- PD explanations are similar to the distilled explanations, both in terms of accuracy and in terms of shape. We clearly see the differences in explanations for the 1H and 2H models.

The behavior on other features and other explanation methods is similar to the one observed with SAT and PD explanations on features $x_4$ and $x_6$.

**Implications:** Since the additive model trained on original labels was able to almost-perfectly match the data (0.02 RMSE), and since all additive models were trained following the same protocols, the differences observed on the distilled explanations can be safely attributed to the behavior of the neural net models. The same conclusion applies to PD, where we see clear differences between the explanations of the 1H and 2H models which are similar to the differences observed on the distilled explanations.

### 5.2 How do different additive explanations characterize non-additive interaction effects?

**Hypothesis:** We focus on $F_2$, introduced in Sect. 4.1. For features $x_3$ to $x_6$, where there are non-zero interactions, we expect the different explanations to differ from the main effects. For features with no interactions ($x_7$ and $x_8$) or where the total effect is the same as the main effect ($x_1$ and $x_2$), we do not expect the explanations to differ significantly from the main effects.

**Setup:** Similar to the previous section, we simulate 50,000 samples and then train a 2H-512,512 neural net to predict $F_2$. We expect the interaction terms of $F_2$ to be absorbed by the different explanation methods. As a comparison, we plot marginal plots that attribute an interaction term to all features that made up that interaction.

**Results:** Fig. 3 shows the main effects (in gray) and total effects (in black) for all the features.

Figure 3 also shows the explanations generated by the different methods. In general, the explanation have relatively similar shapes, although SAT and gSHAP tend to be less smooth than PD. This is not completely surprising, as SAT relies on tree learners, which do not lead to smooth transitions, and gSHAP uses less points than PD to compute the empirical
contribution, since the gSHAP explanation is based on the conditional distribution while PD is based on the marginal.

For features involved in interactions with non-zero expected contributions (i.e., features $x_3$ to $x_6$), explanation methods absorb the contributions as expected. This is particularly clear for features $x_3$, $x_4$, and $x_5$, where the explanations look significantly more similar to the expected contribution of the interactions than to the mains. Out of the explanation methods, gSHAP is the most similar to the expected behavior, but all explanations are relatively similar. 

**Implications:** Additive explanations are able to partially absorb the effect of interactions, as evidenced by the plots of features $x_3$ to $x_6$. However, the allocation of these effects is slightly different for each explanation method. Although we can use total effect as a point of reference, all allocations are valid, and it is not possible to quantify which characterization is better.

### 5.3 How does correlation between features affect additive explanations?

**Hypothesis:** Explanations can be affected by feature correlation in non-obvious ways, for example by one feature absorbing, partially or totally, the importance of a different, correlated feature.

**Setup:** To illustrate the impact of correlation between features on additive explanations, we ran a controlled experiment, modifying the relationship between a feature and the label, but holding the relationship between another correlated feature and the label constant. For this
Table 2 Performance of black-box models. For RMSE, lower is better. For AUC, higher is better

| Data        | n    | p    | Type      | Performance | 1 H  | 2H   |
|-------------|------|------|-----------|-------------|------|------|
| Bikeshare   | 17,000 | 12 | Regression | RMSE        | 0.60 | 0.38 |
| Loan        | 42,506 | 22 | Regression | RMSE        | 2.71 | 1.91 |
| Magic       | 19,000 | 10 | Classification | AUC       | 92.52 | 94.06 |
| Pneumonia   | 14,199 | 46 | Classification | AUC       | 81.81 | 82.18 |
| FICO        | 9,861  | 24 | Classification | AUC       | 79.08 | 79.37 |

Task we use Bikeshare, an UCI data set where the goal is to predict bike usage given 12 features such as “Season”, “Day of the week”, “Temperature”, “Feeling temperature”, or “Humidity”.

For points with temperature between 15 and 18, we doubled the value of the label (the number of rented bikes) without making any modifications to the Feeling temperature feature, which is highly correlated with Temperature (Pearson correlation: 0.988). Separately, for points with humidify between 55 and 65, we increased the value of the label by 1.0 without making any modifications to other features. Humidity has low correlation with other features (the feature with the largest Pearson correlation is Windspeed: 0.346). In both cases, we retrained a 2 H neural net on the modified data, and learned an additive explanation on the modified data. Figure 4 displays the additive explanations trained on the original and modified data. For comparison, we also include the additive explanations for a 1 H neural net trained on the modified data.

Results: Ideally, the additive explanations on the neural net trained on the modified data should be almost identical to the additive explanations on the neural net trained on the original data, except in that particular range of the temperature and humidity features, where we should see an abrupt bump corresponding to the controlled increase that we made. However, we see that a bump also occurs for the Feeling temperature feature, which is highly correlated with temperature.

Implications: When features are correlated, additive explanations of correlated features distributed effects across them. This implies that the analysis of correlated features cannot be done independently. This can become a challenge when the number of correlated features is large.

5.4 How accurate are additive explanations on non-additive models?

Hypothesis: A gap in accuracy between explanations and the original models is expected. Distilled explanations should generally be more accurate than non-distilled explanations.

Setup: We selected five data sets: two UCI data sets (Bikeshare and Magic), a Loan risk scoring data set from an online lending company (Club, 2011), the 2018 FICO Explainable ML Challenge’s credit data set (FICO, 2018), and the pneumonia data set analyzed by Caruana et al. (2015). A 2 H-512,512 neural net is used as the main black-box model. We quantitatively compare the accuracy of different types of global explanations on these data sets: distilled vs non-distilled explanations (e.g. SAT vs gSHAP) and additive vs non-additive explanations (e.g. SAT vs SAT+pairs or decision trees). We also compare the effect of explicitly modeling interactions in the explanation.
Table 3  Performance of global explanations for 2 H black-box models. Performance measured in terms of RMSE for regression tasks and AUROC for classification tasks

| Global explanation | Additive | Distilled | Interactions | Bikeshare RMSE | Loan RMSE | Magic AUC | Pneumonia AUC | FICO AUC |
|--------------------|----------|-----------|--------------|----------------|-----------|-----------|----------------|---------|
| SAT                | Yes      | Yes       | No           | 0.98 ± 0.00    | 2.35 ± 0.01| 90.75 ± 0.06| 82.24 ± 0.05  | 79.42 ± 0.04|
| SAS                | Yes      | Yes       | No           | 0.98 ± 0.00    | 2.34 ± 0.00| 90.58 ± 0.02| 82.12 ± 0.04  | 79.51 ± 0.02|
| Sparse linear      | Yes      | Yes       | No           | 1.39 ± 0.00    | 3.45 ± 0.00| 86.91 ± 0.01| 82.06 ± 0.02  | 79.16 ± 0.01|
| gGRAD              | Yes      | No        | No           | 1.25 ± 0.00    | 6.04 ± 0.01| 80.95 ± 0.13| 81.88 ± 0.05  | 79.28 ± 0.02|
| gSHAP              | Yes      | No        | No           | 1.02 ± 0.00    | 5.10 ± 0.01| 88.98 ± 0.05| 82.31 ± 0.03  | 79.36 ± 0.01|
| PD                 | Yes      | No        | No           | 1.00 ± 0.00    | 4.31 ± 0.00| 82.78 ± 0.00| 82.15 ± 0.00  | 79.47 ± 0.00|
| Decision tree      | No       | Yes       | Yes          | 0.60 ± 0.01    | 2.66 ± 0.02| 91.44 ± 0.29| 79.38 ± 0.38  | 78.19 ± 0.03|
| SAT+pairs          | Yes      | Yes       | Yes          | 0.60 ± 0.00    | 2.13 ± 0.01| 90.75 ± 0.06| 82.23 ± 0.06  | 79.44 ± 0.04|
Fig. 4 Additive explanations of two correlated features before and after label modification experiment. When the labels of certain values of temperature, a feature that is correlated with “Feeling” temperature, are modified, additive explanations of both features, not just the temperature feature, are impacted.

Results: First, for reference, Table 2 presents the accuracy of the 2H-512,512 black-box model, as well as the accuracy of a lower-capacity 1H-8 black-box model (provided for comparison purposes) and additional details about the datasets. Then, Table 3 presents the accuracy of the different explanation methods on these datasets. We draw several conclusions.

- SAT and SAS yield similar accuracy, indicating that the particular choice of the base learner did not matter for these data sets.
- Distilled explanations generally obtain better accuracy than non-distilled explanations. This is not surprising since distilled explanations were trained specifically to mimic the black-box model. However, on most datasets, there is still a gap between the accuracy of the explanations and the accuracy of the black box models.
- Non-linear models such as SAS, SAT, and decision trees outperform the sparse linear models on all datasets. However, sparse linear can be better than decision trees and competitive with SAS and SAT on some datasets (particularly Pneumonia and FICO), suggesting that complex models can easily overfit on some datasets with limited data (FICO and Pneumonia are the datasets with little data and with the largest number of features, cf. Table 2). These datasets are also the ones where the gap between the black-box models and the explanations is the smallest.
- When explicitly modeling interactions, the RMSE in Bikeshare improves from 0.98 to 0.60. The Magic dataset is another example where the interactions brought by the decision tree make a significant difference. However, on some other datasets, decision trees are less accurate than other explanations, and, in the case of the Pneumonia dataset, even less accurate than a simple explanation based on a linear model.
Figure 5 displays explanations of selected features for Magic and Loan. The explanations produced by PD tend to be much too smooth, which hurts its accuracy. Second, in all cases, trees and splines produce similar explanations and obtain equal or better accuracy and fidelity than the other methods. This is not surprising as the other methods are either local methods adapted to the global setting (gSHAP, gGRAD), or are global explanations that are not optimized to learn the teacher’s predictions (PD). For reference, gSHAP when used as a local method (i.e. individual SHAP values, not global explanations) achieved a lower RMSE of 0.37 compared to 1.02 on Bikeshare, and a lower RMSE of 1.99 compared to 5.10 on Loan, which is comparable to its 2H teacher’s RMSE on test data (Table 2). Hence, methods such as gSHAP excel at local explanations and should be used for those, but, to produce global explanations, global model distillation methods optimized to learn the teacher’s predictions perform better.

Implications: Distilled explanations tend to be more accurate than non-distilled explanations. When the goal is to preserve accuracy, global explanations based on distillation are a better choice. Explicitly capturing interactions in the explanations may be unnecessary on some datasets (e.g., FICO) or even pernicious (Pneumonia) while being extremely important in others (e.g., Bikeshare).

6 How interpretable are additive explanations?

Thus far, we have considered how different additive explanations differentially allocate interaction effects, as well as their performance compared to non-additive explanations. We now conduct a user study to evaluate the interpretability of additive explanations compared to non-additive explanations.

Explanations selected. Given the 2H-512,512 neural net described in Sect. 5 on the Bikeshare dataset, we learn four different additive and non-additive explanations using the same distillation techniques: (i) SAT explanations, (ii) Decision tree (DT) explanations, (iii) sparse linear (SPARSE) explanations, and (iv) Subgroup rules (RULES) explanations2. Denoting the “complexity” of the explanations by explanation-K, where the meaning of K depends

---

2 State-of-the-art rule lists (Letham et al., 2015; Angelino et al., 2017) do not support regression, which is needed for distillation. We used a slightly older subgroup discovery algorithm (Atzmueller & Lemmerich, 2012) that supports regression but does not generate disjoint rules. This method only achieved reasonable results on Bikeshare.
on the underlying model,\(^3\) we select the following explanations: SAT-5, DT-4, SAT-2, DT-2, RULES-5, and SPARSE-2.

We selected DT-4 because that is the largest tree that is readable on letter-size paper, and that does not lag too far behind the depth 6 tree in accuracy: the DT-4 has an accuracy of 1.16 RMSE, compared to the 1.0 RMSE of DT-6 and the 0.98 RMSE of SAT. For reference, we show the DT-6 tree in Fig. 8. DT-6 is slightly more accurate than DT-4, but much harder to read. DT-4 used five features: Hour, Temperature, Year, Working Day, Season. To make the explanations more comparable, we select those features for the SAT explanation, hence SAT-5, which has a comparable accuracy of 1.07 RMSE. Figure 6 displays the exact DT-4 and SAT-5 explanations shown to the subjects. We also explored smaller and/or less accurate explanations to assess the trade-off between accuracy and explainability. We selected the top 2 features and built SAT-2 (1.14 RMSE), DT-2 (1.3 RMSE), and SPARSE-2 (1.54 RMSE) explanations. Last, we also considered a rules-based explanation, which is less accurate even with a larger number of features: RULES-5 achieves an RMSE of 1.47, similar to SPARSE-2. Figure 7 displays the exact SPARSE-2 and RULES-5 explanations shown to the subjects.

**Study design.** 50 subjects were recruited to participate in the study. These subjects – STEM PhD students, or college-educated individuals who had taken a machine learning course – were familiar with concepts such as if-then-else structures (for trees and rule lists), reading scatterplots (for SAT), and interpreting equations (for sparse linear models). Each subject only used one explanation model (between-subject design) to answer a set of questions covering common inferential and comprehension tasks on machine learning models: (1) Rank features by importance; (2) Describe relationship between a feature and the prediction; (3) Determine how the prediction changes when a feature changes value; (4) Detect an error in the data, captured by the model. The exact questions were:

(1) What is the most important variable for predicting bike demand?
(2) Rank all the variables from most important to least important for predicting bike demand.
(3) Describe the relationship between the variable Hour and predicted bike demand.
(4) What are variables for which the relationship between the variables and predicted bike demand is positive?
(5) The Hour is 11. When Temperature increases from 15 to 20, how does predicted bike demand change?
(6) There is one error in the data. Any idea where it might be? “Cannot find the error” is an ok answer.

In the first stage, 24 of 50 subjects were randomly assigned to see output from DT-4 or SAT-5. In the second stage, we experimented with smaller versions of trees and SAT using only the two most important features, Hour and Temperature. 14 of 50 subjects were randomly assigned to see output from SAT-2 or DT-2. In the last stage, the remaining 12 subjects were randomly assigned to see output from one of the two worst performing explanations (in terms of accuracy): sparse linear models (SPARSE-2) and subgroup-rules (RULES-5).

**Can humans understand and use explanations?** From the absolute magnitude of the SAT feature explanations as well as Gini feature importance metrics for the tree, we determined the “ground truth” feature importance ranking (in decreasing order): Hour, Temperature, Year, Season, Working Day. More SAT-5 than DT-4 subjects were able to rank the top 2 and all features correctly (75\% vs. 58\%, see Table 4).

---

\(^3\) For decision trees \(K\) represents the depth, and a tree of depth 4 would be denoted as DT-4. For sparse rules, \(K\) represents the number of rules, and a group of 5 rules would be denoted as RULES-5. For SAT and SPARSE, \(K\) denotes the number of features to use. In the case of SPARSE, \(K\) is set indirectly by finding the...
Table 4  Quantitative results from user study. Since SAT-2, DT-2, and SPARSE-2 only had two features, the task to rank five features does not apply. Since the data error only appeared in the output of SAT-5, DT-4, and RULES-5, the other subjects could not have caught the error

| Task                                                      | First stage (n=24) | Second stage (n=14) | Third stage (n=12) |
|-----------------------------------------------------------|--------------------|---------------------|-------------------|
|                                                           | SAT-5              | DT-4                | SAT-2             | DT-2             | SPARSE-2    | RULES-5     |
| Ranked correctly top 2 features                           | 75%                | 58%                 | 100%              | 85.7%            | 83.3%       | 0%          |
| Ranked correctly all (5) features                         | 45%                | 0%                  | N/A               | N/A              | N/A         | 0%          |
| NDCG between human ranking of top 5 features and ground-truth feature importance | 0.94 ± 0.13        | 0.89 ± 0.11         | N/A               | N/A              | N/A         | 0.27 ± 0.11 |
| Described increased demand during rush hour               | 42%                | 0%                  | 29%               | 0%               | 0%          | 33%         |
| Described increased demand during mornings and afternoons  | 33%                | 0%                  | 29%               | 0%               | 0%          | 33%         |
| Compute change in prediction when feature changes         | 33%                | 25%                 | 14%               | 100%             | 83%         | 0%          |
| Caught data error                                         | 33%                | 8%                  | N/A               | N/A              | N/A         | 0%          |
| Time taken (minutes)                                      | 11.7 ± 5.8         | 17.5 ± 14.8         | 7.2 ± 3.2         | 6.2 ± 2.2        | 5.2 ± 3.1   | 14.9 ± 8.4  |
The additive model below was trained to predict bike demand. Each plot describes a variable on the x-axis and predicted bike demand on the y-axis. Hour denotes hour-of-day, i.e. 0, 1, 2, ... 23. Temperature is in Celsius. Year has 2 possible values: 2011, 2012. Season has 4 possible values: Spring, Summer, Fall, Winter.

The tree below was trained to predict bike demand. The tree is using five variables: Hour, Temperature, Year, Season, Working day. Hour denotes hour-of-day, i.e. 0, 1, 2, ... 23. Temperature is in Celsius. Year has 2 possible values: 2011, 2012. Season has 4 possible values: Spring, Summer, Fall, Winter.

When ranking all 5 features, 0% of the DT-4 and RULES-5 subjects were able to predict the right order, while 45% of the SAT-5 subjects correctly predicted the order of the 5 features, showing that ranking feature importance for trees is actually a very hard task. The most common mistake made by DT-4 subjects (42% of subjects) was to invert the ranking of the last two features, Season and Working Day, perhaps because Working Day’s first appearance in the tree (in terms of depth) was before Season’s first appearance (Fig. 6 bottom). We also evaluate the normalized discounted cumulative gain (NDCG) between the ground truth feature importance and the user prediction, where we give relevance scores to the feature regularization lambda parameter that produced the best accuracy on validation while also producing exactly K non-zero feature coefficients.

Fig. 6 Explanations shown to SAT-5 and DT-4 subjects in user study
The lasso (sparse linear regression) model below was trained to predict bike demand. Hour denotes hour-of-day, i.e. 0, 1, 2, … 23. Temperature is in Celsius.

\[
\text{Predicted bike demand} = 1 + 0.056\times\text{Hour} + 0.063\times\text{Temperature}
\]

The rules below were trained to predict bike demand. The rules are not disjoint, but all the rules together make up one machine learning model. If an observation does not fall in any one rule, it gets the last prediction “Otherwise bike demand is …”. Hour denotes hour-of-day, i.e. 0, 1, 2, … 23. Temperature is in Celsius.

\[
\begin{align*}
\text{If year=2012 and workingday=yes} & \quad \text{then bike demand is 2.42} \\
\text{If season=fall and year=2012} & \quad \text{then bike demand is 2.82} \\
\text{If season=summer and year=2012} & \quad \text{then bike demand is 2.61} \\
\text{If season=fall and workingday=yes} & \quad \text{then bike demand is 2.36} \\
\text{If season=fall and year=2012 and workingday=yes} & \quad \text{then bike demand is 2.93} \\
\text{If hour=17 and workingday=yes} & \quad \text{then bike demand is 5} \\
\text{If hour=18 and workingday=yes} & \quad \text{then bike demand is 4.78} \\
\text{If season=summer and yr=2012 and workingday=yes} & \quad \text{then bike demand is 2.57} \\
\text{If hour=08 and workingday=yes} & \quad \text{then bike demand is 4.48} \\
\text{If hour=17 and year=2012} & \quad \text{then bike demand is 5.75} \\
\text{If hour=18 and year=2012} & \quad \text{then bike demand is 5.44} \\
\text{If month=08 and season=fall} & \quad \text{then bike demand is 2.42} \\
\text{Otherwise bike demand is 1.34}
\end{align*}
\]

Fig. 7 Explanations shown to SPARSE-2 and RULES-5 subjects in user study

in decreasing order (i.e., for 5 features, the most important feature has a relevance score of 5, the second most important 4, etc). This gives us an idea of how well the features were ranked, even if the ranking is not perfect. We see how SAT-5 obtains a better score than DT-4, consistent with the previous analysis. RULES-5 obtains a significant lower score.

When asked to describe, in free text, the relationship between the variable Hour and the label, one SAT-5 subject wrote:

\textbf{There are increases in demand during two periods of commuting hours: morning commute (e.g., 7-9 am) and evening commute (e.g., 4-7 pm). Demand is flat during working hours and predicted to be especially low overnight,}

whereas DT-4 subjects’ answers were not as expressive, e.g.:

\textbf{Demand is less for early hours, then goes up until afternoon/evening, then goes down again.}

75% of SAT-5 subjects detected and described the peak patterns in the mornings and late afternoons, and 42% of them explicitly mentioned commuting or rush hour in their description. On the other hand, none of the DT-4 subjects discovered this pattern on the
Fig. 8 Tree of depth 6 (64 leaves), the least deep tree that matched SAT’s fidelity. This uses the default tree visualizer in scikit-learn. Note that the tree is so large that it is hard to read even in digital form at the maximum publishing resolution.
tree: most (58%) described a concave pattern (low and increasing during the night/morning, high in the afternoon, decreasing in the evening) or a positively correlated relation (42%). Similarly, more SAT-5 subjects were able to precisely compute the change in prediction when temperature changed in value, and detect the error in the data – that spring had lower bike demand whereas winter had high bike demand (bottom right feature in Fig. 6 top).

How do tree depth and number of features affect human performance? We also experimented with simpler explanations, SAT-2 and DT-2, that used only the two most important features, Hour and Temperature. As the explanations are simpler, some of the tasks become easier. For example, SAT-2 subjects predict the order of the top 2 features 100% of the time (vs 75% for SAT-5), and DT-2 subjects, 85% of the time (vs 58% for DT-4). The most interesting change is in the percentage of subjects able to compute the change in prediction after changing a feature: only 25% for DT-4, compared to 100% for DT-2. Reducing the complexity of the explanation made using it easier, at the price of reducing the accuracy of the explanation. Another important aspect is the time needed to perform the tasks: increasing the number of features from 2 to 5 increases the time needed by the subjects to finish the study by 60% for the SAT explanation, but increases it by 166% for the DT explanation, that is, interpreting a tree becomes much more costly as the tree becomes deeper (and more accurate), and, in general, subjects make more mistakes. SAT scales up more gracefully.

Remaining explanations: subgroup-rules and sparse linear models. These explanations were the least accurate and faithful. We found that human subjects can easily read the (few) weights of SPARSE-2, establish feature importance, and compute prediction changes, and do so quickly – at 5.1 minutes on average, this was the fastest explanation to interpret. However,
the explanation is highly constrained and hid interesting patterns. For example, 100% of the subjects described the relation between demand and hour as increasing, and 83% predicted the exact amount of increase, but none were able to provide insights like the ones provided by SAT-5 and DT-4 subjects.

RULES-5 was the second hardest explanation to interpret based on mean time required to answer the questions: 14.9 min. Understanding non-disjoint rules appears to be hard: none of the subjects correctly predicted the feature importance order, even for just two features; none were able to compute exactly the change in prediction when feature value changes, and none were able to find the data error. The rules in RULES-5 are not disjoint because we could not find a regression implementation of disjoint rules. However, 66% of the subjects discovered the peak during rush hour, as that appeared explicitly in some rules, e.g. “If hour=17 and workingday=yes then bike demand is 5”.

**Fidelity vs. interpretability.** Fig. 9 presents detailed results for individual users by model. On the left is the time needed to finish the study (left). In the center is the length of the user’s written description of the relationship between a feature and model predictions. On the right is the NDCG rank loss of user ranking of feature importance compared to ground-truth feature importance. All of these metrics can be considered interpretability metrics, when defining interpretability as grounded in human tasks (Doshi-Velez & Kim, 2018). On the y-axis is fidelity (RMSE), i.e., how similar is the explanation to the 2 H model prediction.

The plots show that there is a trade-off between fidelity and interpretability (as measured by time to complete, description length, and NDCG of feature rankings), but not all methods behave similarly. In general, the SPARSE-2 explanation is easy to understand (users typically finish the study rapidly), but fidelity is poor and it leads to short descriptions. On the other hand, SAT-5 and DT-4 have much better fidelity and lead to more varied descriptions, but also took longer to understand. DT-2 was faster to complete than DT-4, but the fidelity is lower and the descriptions shorter. RULES-5 is better than SPARSE-2, but not as good as SAT-5 or DT-4. SAT-5 offers a reasonable trade-off, being both faithful and relatively easy to understand, while also leading to rich descriptions for many users.

To summarize, global additive explanations: (1) allowed humans to perform better (than decision trees, sparse linear models, and rules) at ranking feature importance, pointing out patterns between certain feature values and predictions, and catching a data error; (2) Additive explanations were also faster to understand than big decision trees; (3) Very small decision trees and sparse linear models had the edge in calculating how predictions change when feature values change, but were much less faithful and accurate.

### 7 Conclusions

In this work we studied how additive explanations behave when explaining black-box models with non-additive components. We showed that in this case, there is no one unique additive explanation, with different explanation methods characterize these non-additive components in different ways. We quantitatively compared different additive explanations on several regression and classification tasks, finding that distilled explanations are generally the most accurate additive explanations. Although non-additive explanations that explicitly model interactions tend to be more accurate, we found, through a user study, that machine learning practitioners were able to leverage additive explanations better. When deciding which explanation to use for a black-box model, trustworthiness, accuracy, and interpretability should all be considered.
Acknowledgements  We thank Julius Adebayo for helpful discussion.

Author contributions  The authors contributed to this paper in the following manner: Sarah Tan designed, executed, and analyzed the experiments and wrote the paper. Giles Hooker formulated mathematics in the paper and wrote the paper. Paul Koch wrote software used by the experiments. Albert Gordo executed experiments and wrote the paper. Rich Caruana analyzed experiments and wrote the paper.

Funding  Giles Hooker was supported by NSF Grant No DMS-1712554.

Data availability  Publicly available datasets used in this paper can be found at the websites of: UCI (Bikeshare, Magic), Kaggle (Lending Club), FICO. Whenever available, links are cited in this paper’s references.

Code Availability  Publicly-available code to help other researchers replicate the work can be found at: https://github.com/shftan/distilled_additive_explanations.

Declarations

Conflict of interest  The authors declare that there is no conflict of interest.

Consent to participate  User study subjects consented to participate in the user study.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2019). Sanity checks for saliency maps. In NeurIPS.
Amodio, S., Aria, M., & D’Ambrosio, A. (2014). 2014. Statistica: On concurrence in nonlinear and nonparametric regression models.
Ancona, M., Ceolini, E., OztiReki, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In ICLR.
Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists. In KDD.
Apley, D.W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82, 4.
Atzmueller, M., & Lemmerich, F. (2012). VIKAMINE - Open-source subgroup discovery, pattern mining. In ECML PKDD: Analytics.
Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep?. In NeurIPS.
Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS ONE, 10, 7.
Bastani, O., Kim, C., & Bastani, H. (2017). Interpreting blackbox models via model extraction. In FAT/ML Workshop.
Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Quantifying interpretability of deep visual representations: Network dissection. In CVPR.
Bhatt, U., Weller, A., & Moura, J. M. F. (2020). Evaluating and aggregating feature-based model explanations. In IJCAI.
Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. The Annals of Applied Statistics, 5, 4.
Breiman, L. (2001). Random forests. Machine Learning, 45, 1.
Bucilua, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In KDD.
Caruana, R., Lou, Y., Gehreke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In KDD.
Chang, C.H., Tan, S., Lengerich, B., Goldenberg, A., & Caruana, R. (2021). How interpretable and trustworthy are GAMs. In KDD.
Covert, I., Lundberg, S., & Lee, S.I. (2020). Understanding global feature contributions through additive importance measures. In NeurIPS.
Craven, M. W., & Shavlik, J. W. (1995). Extracting tree-structured representations of trained networks. In NeurIPS.
Doshi-Velez, F., & Kim, B. (2018). Towards A rigorous science of interpretable machine learning. In Explainable and interpretable models in computer vision and machine learning. Springer.
FICO. (2018). FICO explainable machine learning challenge. https://community.fico.com/s/explainable-machine-learning-challenge.

Fisher A.J., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. JMLR.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29, 5.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. The Annals of Applied Statistics, 2, 3.

Frosst, N., & Hinton, G. (2018). Distilling a neural network into a soft decision tree. In CEUR-WS.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In AISTATS.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. Journal of Statistical Science, 1(3), 297–310.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.

Hinton, G., Vinyals, O., & Dean, J. (2014). Distilling the knowledge in a neural network. In NeurIPS Deep learning and representation learning workshop.

Ibrahim, M., Louie, M., Modarres, C., & Paisley, J. (2019). Mapping the landscape of predictions: Global explanations of neural networks. In AIES.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML.

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). How can I choose an explainer? In F AccT: An application-grounded evaluation of post-hoc explanations.

Kim, B., Khanna, R., & Koyejo, O. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In ICML.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In NeurIPS.

Leducich, B., Tan, S., Chang, C. H., Hooker, G., & Caruana, R. (2020). An efficient algorithm for recovering identifiable models: Purifying interaction effects with the functional anova. In AISTATS.

Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics, 9, 3.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NeurIPS.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NeurIPS.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NeurIPS.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NeurIPS.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NeurIPS.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NeurIPS.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NeurIPS.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NeurIPS.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NeurIPS.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
Sanchez, I., Rocktaschel, T., Riedel, S., & Singh, S. (2015). Towards extracting faithful and descriptive representations of latent variable models. In *AAAI spring symposium on knowledge representation and reasoning: Integrating symbolic and neural approaches*.
Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GIlocalX-from local to global explanations of black box AI models. *Artificial Intelligence* 294.
Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *ICML*.
Sanchez, I., Rocktaschel, T., Riedel, S., & Singh, S. (2015). Towards extracting faithful and descriptive representations of latent variable models. In *AAAI spring symposium on knowledge representation and reasoning: Integrating symbolic and neural approaches*.
Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX-from local to global explanations of black box AI models. *Artificial Intelligence* 294.
Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *ICML*.
Simonyan, K., & Vedaldi, A., Zisserman, A. (2014). Visualising image classification models and saliency maps. In *ICLR Workshop: Deep inside convolutional networks*.
Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*.
Sobol’, I. M. (1990). On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2, 1.
Slack, D., Hilgard, S., Jia, E., & Singh, S., Lakkaraju, H. (2020). Adversarial attacks on post hoc explanation methods: Fooling LIME and SHAP. In *AIES*.
Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41, 3.
Tan, S. (2018). Interpretable approaches to detect bias in black-box models. In *AIES doctoral consortium*.
Tan, S., Caruana, R., & Hooker, G., Lou, Y. (2018). Auditing black-box models using transparent model distillation: Distill-and-compare. In *AIES*.
Tan, S., Soloviev, M., Hooker, G., & Wells, M. T. (2020). Tree space prototypes: Another look at making tree ensembles interpretable. In *FODS*.
Tsang, M., Cheng, D., & Liu, Y. (2018). Detecting statistical interactions from neural network weights. In *ICLR*.
vander Linden, I., Haned, H., & Kanoulas, E. (2019). Global aggregations of local explanations for black box models. In *SIGIR Fairness, accountability, confidentiality, transparency, and safety workshop*.
Williamson, B., & Feng, J. (2020). Efficient nonparametric statistical inference on population feature importance using Shapley values. In *ICML*.
Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Chapman and Hall/CRC.
Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society: Series B*, 73, 1.
Yan, T., & Procaccia, A. D. (2021). If you like shapley then you’ll love the core. In *AAAI*.
Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39, 1.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.