

Considerations When Learning Additive Explanations for Black-Box Models

SARAH TAN, Facebook

GILES HOOKER, University of California, Berkeley

PAUL KOCH, Microsoft Research

ALBERT GORDO, Facebook

RICH CARUANA, Microsoft Research

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.

Additional Key Words and Phrases: 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 [45], a local feature importance method, consists of Shapley values that add up to the prediction for a particular point. LIME [53] fits a locally linear 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 the prediction function [28]. Even when models only have additive components, but different components are correlated [2], there is no one unique additive explanation. This has been demonstrated in several practical settings; Orlenko and Moore [49] 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. [41] 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?*

* This work started before the authors joined Facebook.

Authors' addresses: Sarah Tan, Facebook, sarahtan@fb.com; Giles Hooker, University of California, Berkeley; Paul Koch, Microsoft Research; Albert Gordo, Facebook; Rich Caruana, Microsoft Research.

We study this question from several angles in this paper. First, we show how several popular additive explanation methods – partial dependence [23], Shapley additive explanations [45], distilled additive explanations [63], 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 [29]), 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.

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 remaining of the paper is as follows. Section 2 discusses related work. In Section 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 Section 4, and provide results of empirical investigations of these phenomena in Section 5. In Section 6 we describe results from the user study. Section 7 concludes the paper.

2 RELATED WORK

Global explanations. Post-hoc explanations of black-box models can be roughly divided into local methods that explain predictions of individual points and global methods that explain overall model predictions [20]. Examples of global explanations include using a second interpretable model to approximate the black-box model [9, 19, 25, 26, 32, 40, 55], analyzing intermediate representations [10, 47] or concepts [37] encoded by the black-box model, prototype selection [12, 36, 65], counterfactual explanations [52] that summarize actions that can be taken to change the black-box model’s predictions for an entire population, and feature importance measures [13, 18, 22, 45, 68, 71]. 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.

Trustworthiness of explanations. One viewpoint in the community is that post-hoc explanations of black box models can be unreliable or misleading [54], 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 [60]. 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 [41]. In addition, different training strategies may lead to significantly different or even contradictory models [16]. In this paper we empirically investigate this phenomenon on additive explanations.

Evaluation of interpretability. There is no universal definition of interpretability [20]; 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) [39, 51]; recently, there has been more work evaluating interpretability on expert users [34, 35, 40]. 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 by p matrix consisting of feature values of points $\{x_i^j, i = 1, \dots, p, j = 1, \dots, n\}$ where p is the number of features and n is the number of points. Let the vector $X^j = (x_1^j, \dots, x_p^j)$ represent all feature values of the j th point, and let the vector $X_i = (x_i^1, \dots, x_i^n)$ 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:

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

where $\|F(\mathbf{X}) - G(\mathbf{X})\| = \text{error}_G$. $h_i(X_i)$ 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. This decomposition is not unique and each of the methods examined below will give different decompositions. 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 $X^j = (x_1^j, \dots, x_p^j)$, local additive explanations Γ^j decompose $F(X^j)$, as a sum of effects of p features:

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

In Section 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 Figure 1 – 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 $h_i(x_i)$. For example, partial dependence plots [23] and Shapley dependence plots [45] 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. For feature X_i , partial dependence [23] 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, partial dependence plots are constructed by repeatedly querying the prediction function of the black-box model with new points generated by setting the value of feature X_i to a specific value z :

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

Partial dependence plots of different features can be summed up as in Equation 1 with $h_i(X_i) = q_i(X_i)$ to construct a global additive explanation. However, since each $q_i(X_i)$ has the same average value as the prediction function F , the sum $\sum_{i=1}^p q_i(X_i)$ is no longer on the scale of F . This can be remedied by centering each $q_i(X_i)$ by subtracting

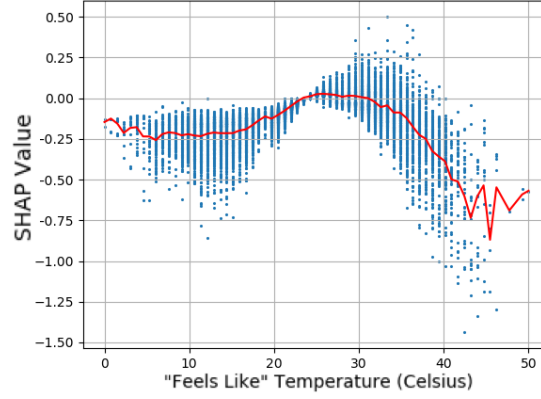


Fig. 1. From [local Shapley explanations](#) to [gSHAP](#), a global Shapley additive explanation for a particular feature created by aggregating local Shapley explanations.

a constant before performing the summation. A natural constant we consider is $\overline{q_i(X_i)} = \frac{1}{|z|} \sum_z \widehat{q_i}(X_i = z)$, the average of $q_i(X_i)$ over all possible values of z , to obtain the global additive explanation:

$$PD(\mathbf{X}) = \hat{h}_0 + \sum_{i=1}^p \left\{ \widehat{q_i}(X_i) - \overline{q_i(X_i)} \right\} \quad (3)$$

While partial dependence takes the expectation of the prediction function over the *marginal* distribution of all features besides X_i , this expectation could also be taken over the *conditional* distribution of all features besides X_i , conditional on X_i having value z . This is the marginal plot¹ [5, 29, 72], 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 summing up predictions of existing points, $N(z)$, whose value of feature X_i is a specific value z :

$$\widehat{r_i}(X_i = z) = \frac{1}{|N(z)|} \sum_{j \in N(z)} F(x_1^j, \dots, x_{i-1}^j, x_i^j = z, x_{i+1}^j, \dots, x_p^j | x_i^j = z)$$

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

$$M(\mathbf{X}) = \hat{h}_0 + \sum_{i=1}^p \left\{ \widehat{r_i}(X_i) - \overline{r_i(X_i)} \right\} \quad (4)$$

where $\overline{r_i(X_i)}$ is constructed using a similar centering procedure on $r_i(X_i)$ 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.

¹The name is counter-intuitive as it is based on the conditional, not marginal distribution.

3.2.2 Shapley Additive Explanations. The Shapley local additive explanation [45] for a point X^j consists of ϕ_i^j that sum up as in Equation 2 to $F(X^j)$, the prediction for point X^j , where:

$$\phi_i^j(x_i^j) = \sum_{S \subseteq \{x_1^j, \dots, x_p^j\} \setminus x_i^j} \frac{|S|!(p - |S| - 1)!}{p!} [F(S \cup \{x_i^j\}) - F(S)] \quad (5)$$

with S being some subset of features excluding the i th feature. For example, each blue point in Figure 1 is one such ϕ_i^j for feature $i = \text{“Temperature”}$ and a particular point j . In practice, Equation 5 is computed by approximation, using techniques that repeatedly query the black-box model such as Kernel SHAP [45], Shapley sampling values [62], etc.

Local explanations that act on single points can also be aggregated to construct a global explanation [11, 56, 67]. For example, the red line in Figure 1 was created using the basic approach of feature-wise sample mean [11], averaging $\phi_{\text{Temperature}}^j$ at each unique temperature value z to construct a global additive explanation that we call **gSHAP** and define as:

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

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 [18] 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 [7, 14, 30] 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 [19, 25], generalized additive models [63], linear models, etc. If an additive model is used, the resulting explanation can be written as:

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

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

$$\begin{aligned} \hat{h}_0, \hat{h}_1, \dots, \hat{h}_p &= \operatorname{argmin} L(h_0, h_1, \dots, h_p) = \operatorname{argmin} \frac{1}{n} \sum_{j=1}^n \|F(X^j) - G(X^j)\|_2^2 \\ &= \operatorname{argmin} \frac{1}{n} \sum_{j=1}^n \|F(X^j) - (h_0 + \sum_{i=1}^p h_i(x_i^j))\|_2^2 \end{aligned}$$

In this paper we consider two types of additive models: spline-based [69] and tree-based [43] 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 [70]. For SAT, $\hat{h}_i(X_i)$ is a bagged tree function of feature X_i implemented using the InterpretML [48] package. Pairwise interactions can also be modeled explicitly [44], in the form of $h_{i,i'}(x_i^j, x_{i'}^j)$ added to Equation 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 $gGRAD$ through the first-order Taylor decomposition of F :

$$gGRAD(\mathbf{X}) = F(\mathbf{0}) + \sum_{i=1}^p \frac{1}{n} \sum_{j=1}^n \frac{\partial F(X^j)}{\partial x_i^j} x_i^j \quad (8)$$

This is related to the “gradient*input” method (e.g. [57]) used to generate saliency maps for images. While there are several other gradient-based feature attribution methods ([8, 57–59]; also see [46] or [3] for a review), the “gradient*input” method is simple, intuitive, and studied a lot in the literature [1, 3], 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:

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

by decomposing F into successive additive components of increasing complexity and using the error term to capture any remaining effects.

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 :

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

For example, when func is the variance function, this is the main effect Sobol index [50, 61]. 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 *all* features that were 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 :

$$T(X_i) = \text{func}\{e_i(X_i) + \sum_{i'=1, i \neq i'}^p e_{ii'}(X_i, X_{i'})\} \quad T(X_{i'}) = \text{func}\{e_{i'}(X_{i'}) + \sum_{i'=1, i \neq i'}^p e_{ii'}(X_i, X_{i'})\} \quad (10)$$

For example, when func is the variance function, this is the total effect Sobol index [50, 61], and when 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 [29], also called marginal plots in [5]) or partial dependence plots [23]). In cases where 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 different features – For example, distillation approaches that train additive student models to mimic the prediction functions of non-additive black-box models [64] express non-additive components as a best-fit additive approximation added to additive components. Shapley additive explanations [45] 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:

$$F_2(\mathbf{x}) = F_1(\mathbf{x}) + x_1x_2 + |x_3|^{2|x_4|} + \sec(x_3x_5x_6) \quad (11)$$

where

$$F_1(\mathbf{x}) = 3x_1 + x_2^3 - \pi^{x_3} + \exp(-2x_4^2) + \frac{1}{2 + |x_5|} + x_6 \log(|x_6|) + \sqrt{2|x_7|} + \max(0, x_7) + x_8^4 + 2 \cos(\pi x_8) \quad (12)$$

This non-additive, highly nonlinear function consists of components from synthetic functions proposed by [24], [31] and [66]. Like [66], we set the domain of all features to be independent and Uniform[-1,1]. Like [24], 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 Figure 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 $E[x_1x_2|x_2] = E[x_1x_2|x_1] = \int_{-1}^1 x_1x_2 dP(x_2|x_1) = \frac{x_1}{2} \frac{x_2^2}{2} \Big|_{-1}^1 = 0$, despite the presence of an interaction term x_1x_2 in F_2 . As we will see in Figure 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 fully-connected neural nets (FNNs) with ReLU nonlinearities as black box models, though other models such as gradient boosted trees or random forests could also be used. Their prediction functions F typically contain non-additive interaction effects. The prediction function is obtained by having the black-box model label a set of training data. Across datasets used in this paper, the most accurate neural nets were FNNs with 2-hidden layers and 512 hidden units per layer (2H-512,512); nets with three or more hidden layers had a lower training loss, but did not generalize well on most problems. In some experiments we also used a restricted-capacity neural net with 1 hidden layer of 8 units (1H-8). Training was performed in PyTorch. We use the Adam optimizer [38] with default beta parameters, Xavier initialization [27], and early stopping based on validation loss. For each depth, we use random search to find the optimal hyperparameters (number of hidden units, learning rate, weight decay, dropout probability, batch size, enabling batch norm [33], etc) based on average validation performance on multiple train-validation splits and multiple random initializations.

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

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.

Quantitative evaluation of explanations: We will quantitatively evaluate explanations as if they were models, following Lundberg and Lee [45] 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 [36] who used their explanations (prototypes) to classify test data.

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 Section 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: Figure 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

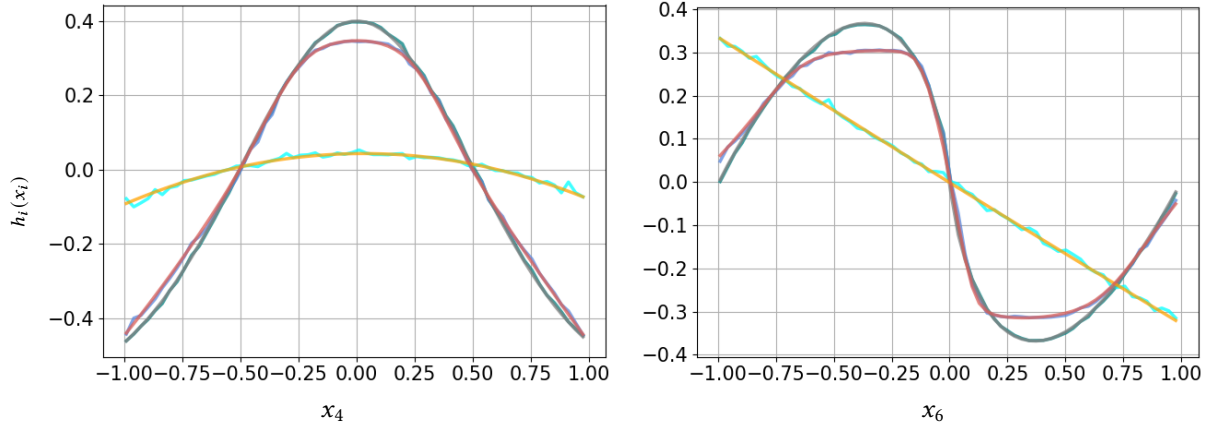


Fig. 2. For two features of function F_1 : **Main effect**, **GAM trained on data**, **SAT explanation of complex black-box model (2H-512,512 neural net)**, **PD explanation of the same 2H model**, **SAT of simpler black-box model (1H-8 neural net)**, and **PD explanation of the same 1H model**.

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.

- 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 Section 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: Figure 3 shows the main effects (in gray) and total effects (in black) for all the features.

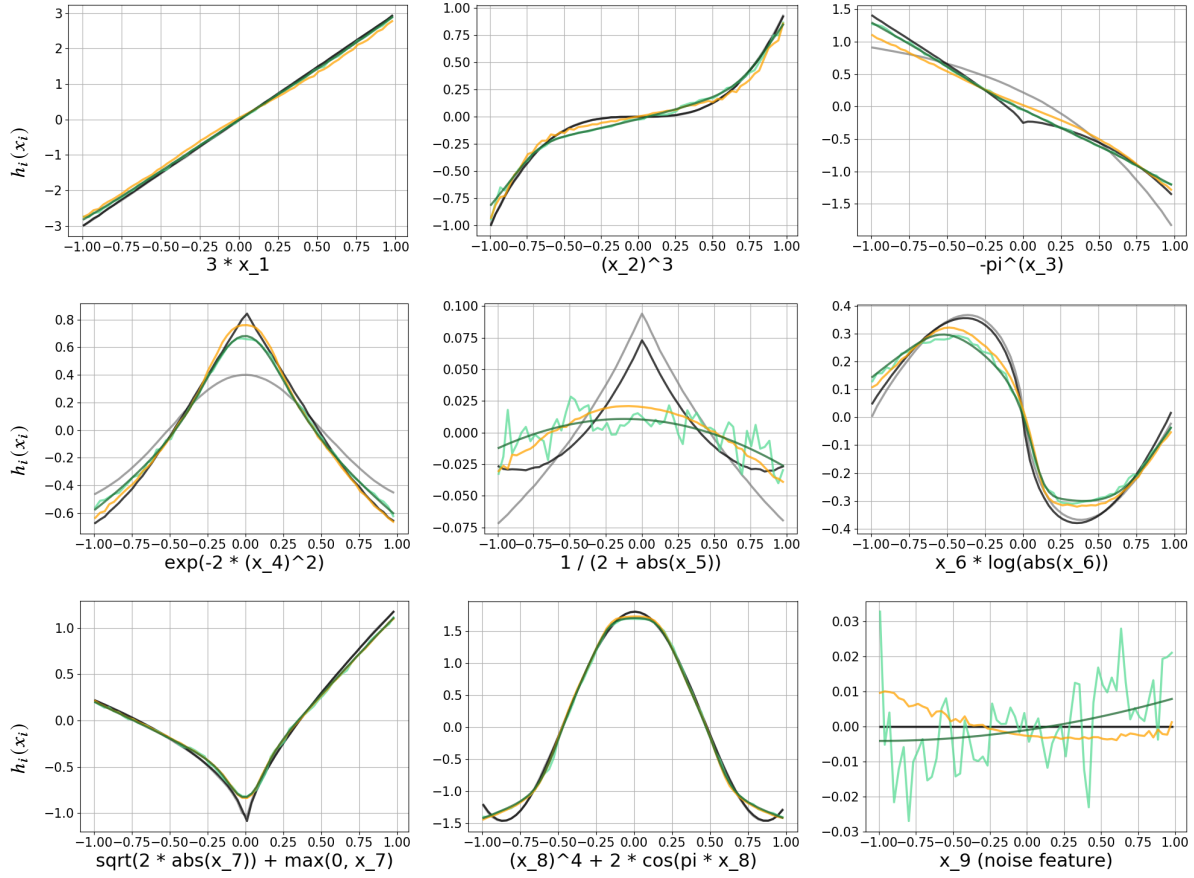


Fig. 3. For all features of function F_2 : Main effect, **Total effect**, SAT explanation of complex black-box model (2H-512,512 neural net), PD explanation of the same 2H model, and gSHAP explanation of the same 2H model. Note that not all features are in the same scale, and that the scale of features x_5 and x_9 (a noise feature) are significantly smaller than the remaining 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

Data	n	p	Type	Performance		
					1H	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

Table 2. Performance of black-box models. For RMSE, lower is better. For AUC, higher is better.

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 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 humidity 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 2H 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 1H 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.

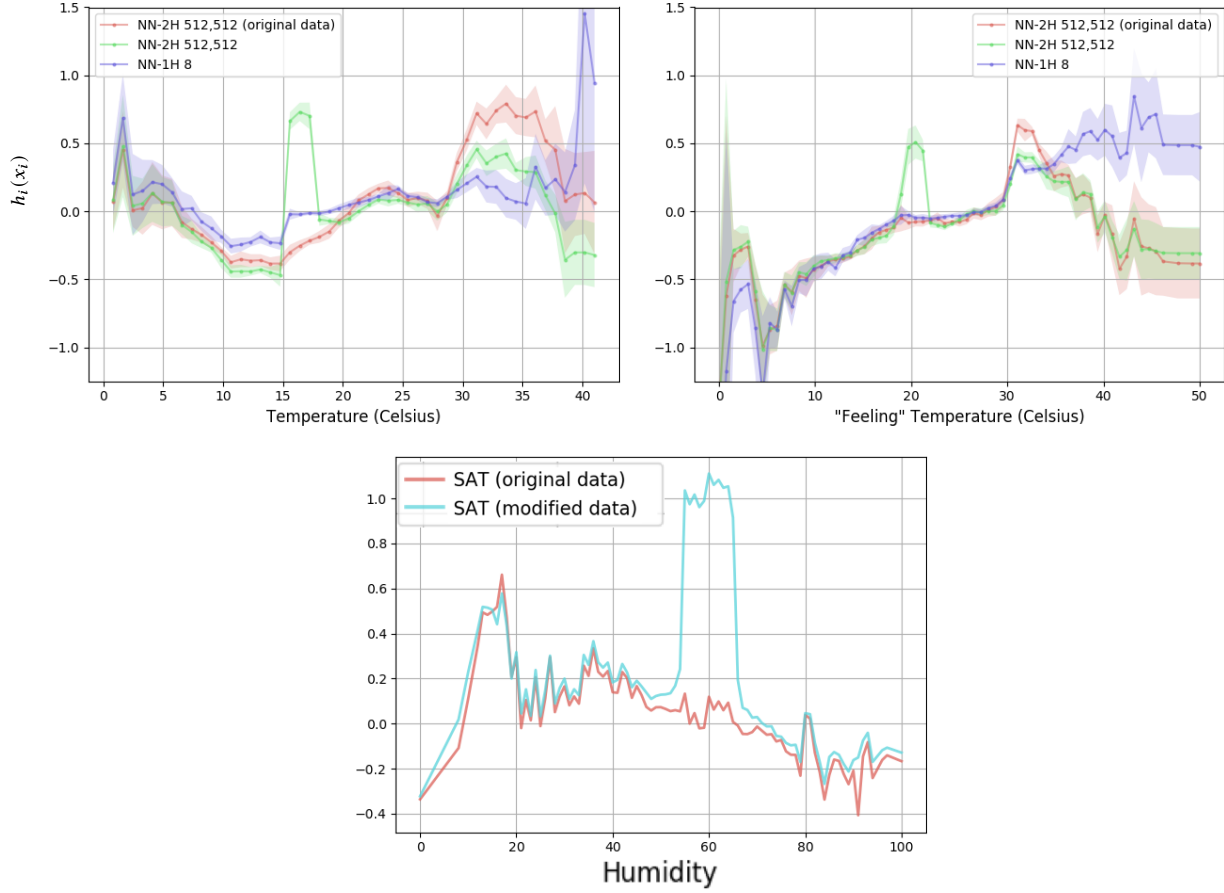


Fig. 4. Additive explanations of two correlated features before and after label modification experiment. When the labels of certain values of temperature, a feature that is correlated with “Feeling” temperature, are modified, additive explanations of **both** features, not just the temperature feature, are impacted

Setup: We selected five data sets: two UCI data sets (Bikeshare and Magic), a Loan risk scoring data set from an online lending company [17], the 2018 FICO Explainable ML Challenge’s credit data set [21], and the pneumonia data set analyzed by [15]. A 2H-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.

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.

Accuracy				Bikeshare	Loan	Magic	Pneumonia	FICO
Global Explanation	Additive	Distilled	Interactions	RMSE	RMSE	AUC	AUC	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

Table 3. Performance of global explanations for 2H black-box models. Performance measured in terms of RMSE for regression tasks and AUROC for classification tasks.

- 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).

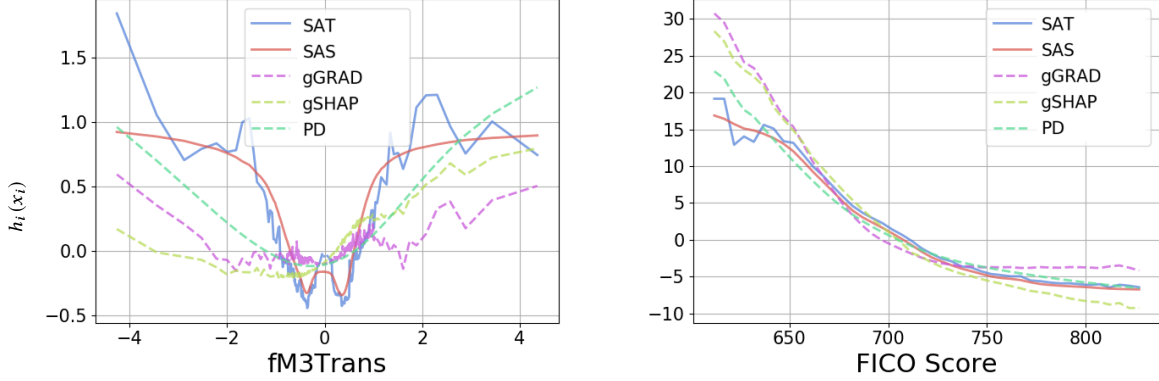


Fig. 5. Example explanations of selected features for Magic data (left), and Loan data (right). SAT and SAS tend to agree.

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 Section 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 on the underlying model (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) 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 Figure 8. 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

²State-of-the-art rule lists [4, 42] do not support regression, which is needed for distillation. We used a slightly older subgroup discovery algorithm [6] that supports regression but does not generate disjoint rules. This method only achieved reasonable results on Bikeshare.

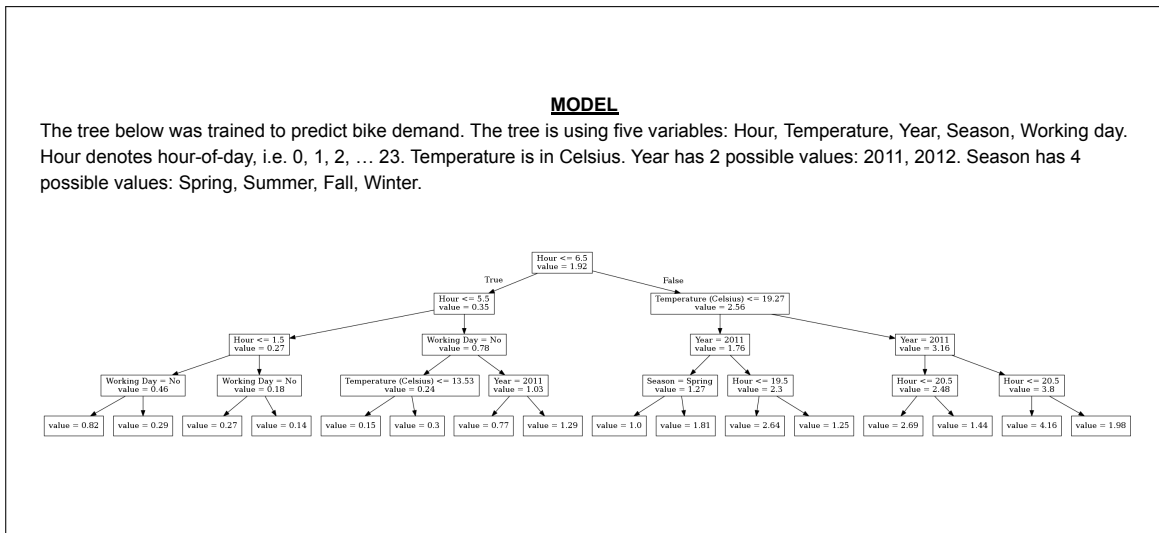
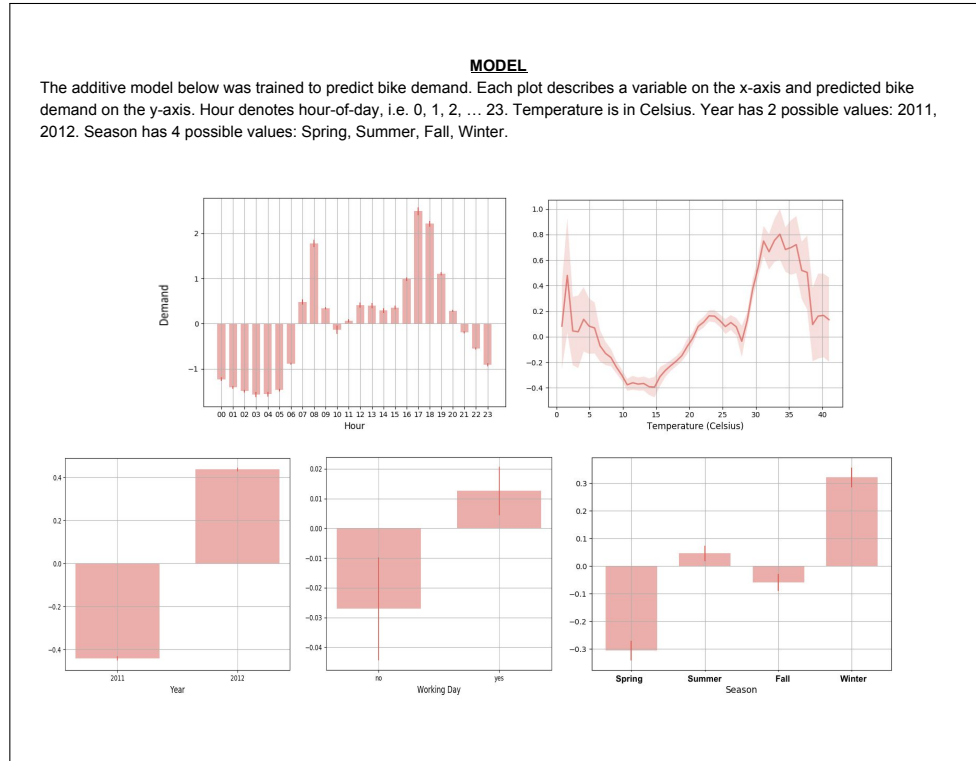


Fig. 6. Explanations shown to SAT-5 and DT-4 subjects in user study

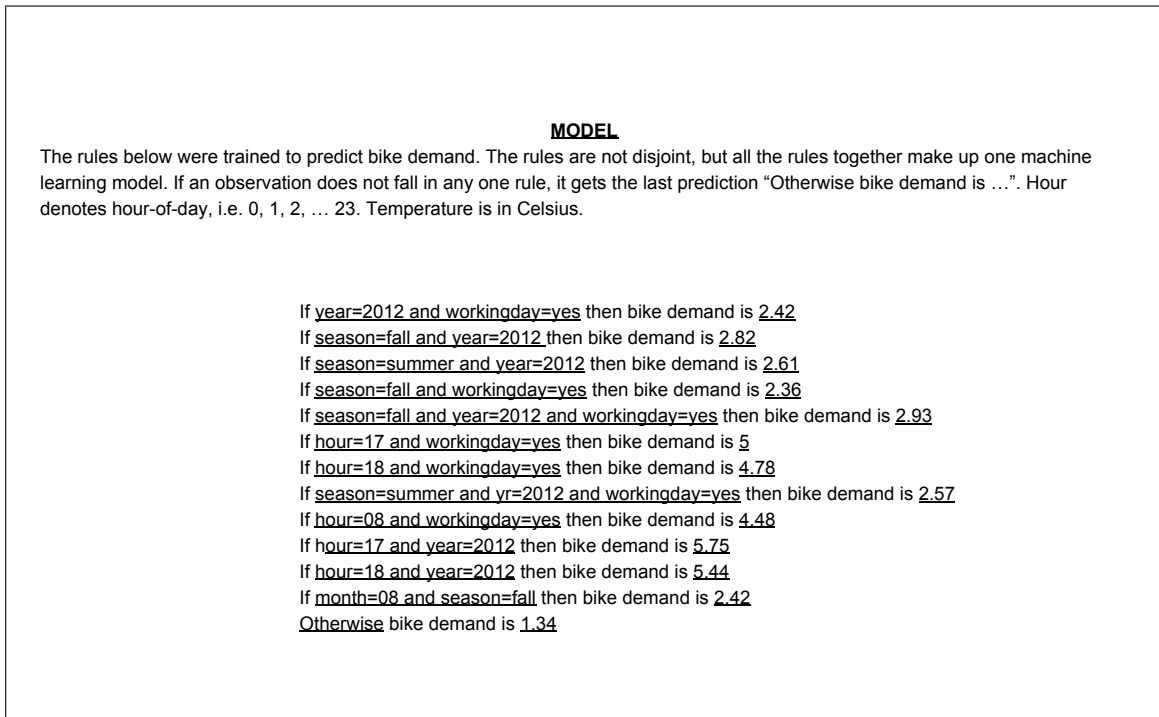
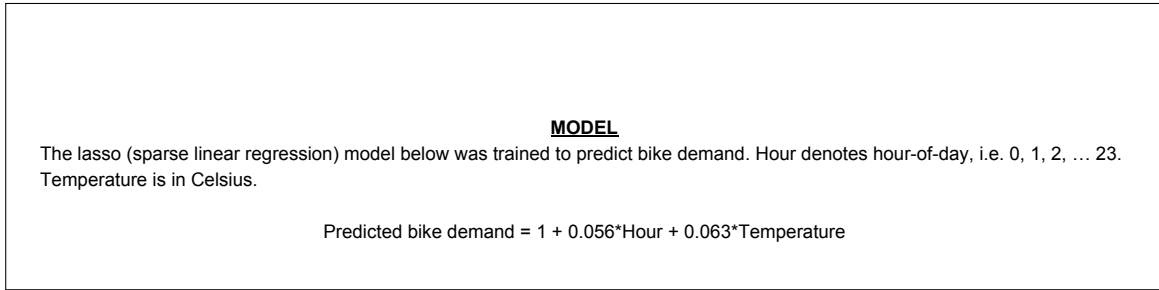


Fig. 7. Explanations shown to SPARSE-2 and RULES-5 subjects in user study

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.

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 \pm 0.13	0.89 \pm 0.11	N/A	N/A	N/A	0.27 \pm 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 \pm 5.8	17.5 \pm 14.8	7.2 \pm 3.2	6.2 \pm 2.2	5.2 \pm 3.1	14.9 \pm 8.4

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.

- (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).

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 (Figure 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 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.



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.

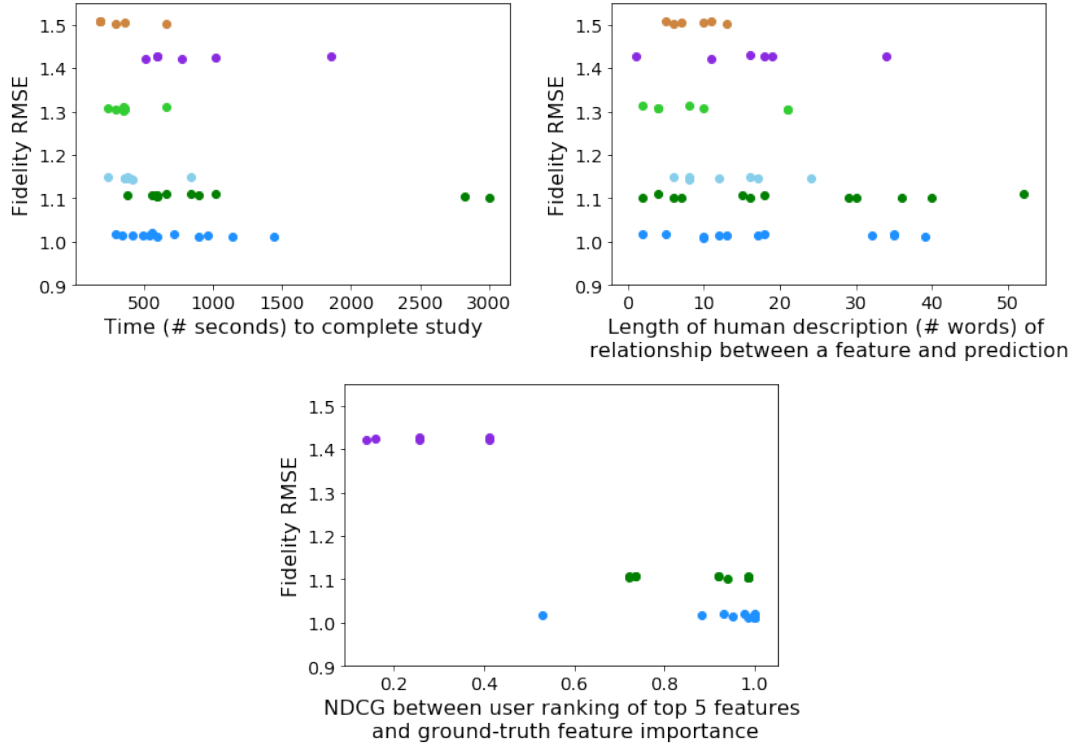


Fig. 9. User study metrics, as proxies for interpretability, by fidelity (RMSE) for different explanations. Each point is an individual user in the user study. The metrics are time needed to finish the study (top left), length of the description (top right), and the NDCG of the ranked features (bottom). Key: SAT-5, DT-4, SAT-2, DT-2, RULES-5, SPARSE-2

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

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.:

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 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 Figure 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 minutes. 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. Figure 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 [20]. On the y-axis is fidelity (RMSE), *i.e.*, how similar is the explanation to the 2H 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.

8 ACKNOWLEDGEMENTS

We thank Julius Adebayo for helpful discussion. GH was supported by NSF grant DMS-1712554.

REFERENCES

- [1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2019. Sanity checks for saliency maps. In *NeurIPS*.
- [2] Sonia Amodio, Massimo Aria, and Antonio D’Ambrosio. 2014. On concavity in nonlinear and nonparametric regression models. *Statistica* (2014).
- [3] Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *ICLR*.
- [4] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. 2017. Learning Certifiably Optimal Rule Lists. In *KDD*.
- [5] Daniel W Apley and Jingyu Zhu. 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 (2020).
- [6] Martin Atzmueller and Florian Lemmerich. 2012. VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In *ECML PKDD*.
- [7] Jimmy Ba and Rich Caruana. 2014. Do Deep Nets Really Need to be Deep?. In *NeurIPS*.
- [8] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Muller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10, 7 (2015).
- [9] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpreting blackbox models via model extraction. In *FAT/ML Workshop*.
- [10] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *CVPR*.
- [11] Umang Bhatt, Adrian Weller, and José M. F. Moura. 2020. Evaluating and Aggregating Feature-based Model Explanations. In *IJCAI*.
- [12] Jacob Bien and Robert Tibshirani. 2011. Prototype selection for interpretable classification. *The Annals of Applied Statistics* 5, 4 (2011).
- [13] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001).
- [14] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *KDD*.
- [15] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*.
- [16] Chun-Hao Chang, Sarah Tan, Ben Lengerich, Anna Goldenberg, and Rich Caruana. 2021. How Interpretable and Trustworthy are GAMs. In *KDD*.
- [17] Lending Club. 2011. Lending Club Loan Dataset 2007-2011. <https://www.lendingclub.com/info/download-data.action>.
- [18] Ian Covert, Scott Lundberg, and Su-In Lee. 2020. Understanding global feature contributions through additive importance measures. In *NeurIPS*.
- [19] Mark W. Craven and Jude W. Shavlik. 1995. Extracting Tree-structured Representations of Trained Networks. In *NeurIPS*.
- [20] Finale Doshi-Velez and Been Kim. 2018. Towards A Rigorous Science of Interpretable Machine Learning. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer.
- [21] FICO. 2018. FICO Explainable Machine Learning Challenge. <https://community.fico.com/s/explainable-machine-learning-challenge>.
- [22] Aaron J. Fisher, C. Rudin, and F. Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously. *JMLR* (2019).
- [23] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29, 5 (2001).
- [24] Jerome H. Friedman and Bogdan E. Popescu. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics* 2, 3 (2008).

- [25] Nicholas Frosst and Geoffrey Hinton. 2018. Distilling a Neural Network Into a Soft Decision Tree. In *CEUR-WS*.
- [26] LiMin Fu. 1994. Rule generation from neural networks. *IEEE Transactions on Systems, Man, and Cybernetics* 24, 8 (1994).
- [27] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*.
- [28] Trevor Hastie and Rob Tibshirani. 1986. Generalized Additive Models. *Statist. Sci.* (1986).
- [29] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.
- [30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the Knowledge in a Neural Network. In *NeurIPS Deep Learning and Representation Learning Workshop*.
- [31] Giles Hooker. 2004. Discovering additive structure in black box functions. In *KDD*.
- [32] Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. 2019. Global Explanations of Neural Networks: Mapping the Landscape of Predictions. In *AIES*.
- [33] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*.
- [34] Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. 2021. How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations. In *FAccT*.
- [35] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *CHI*.
- [36] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *NeurIPS*.
- [37] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Jun Cai, James Wexler, Fernanda Viegas, and Rory Abbott Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *ICML*.
- [38] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.
- [39] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. In *HCOMP*.
- [40] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In *AIES*.
- [41] Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. 2020. Purifying interaction effects with the functional anova: An efficient algorithm for recovering identifiable additive models. In *AISTATS*.
- [42] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3 (2015).
- [43] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *KDD*.
- [44] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *KDD*.
- [45] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *NeurIPS*.
- [46] Gregoire Montavon, Wojciech Samek, and Klaus-Robert Muller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73 (2018).
- [47] Jesse Mu and Jacob Andreas. 2020. Compositional Explanations of Neurons. In *NeurIPS*.
- [48] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* (2019).
- [49] Alena Orlenko and Jason H Moore. 2021. A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions. *BioData Mining* 14, 1 (2021).
- [50] Art B Owen. 2014. Sobol' indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification* (2014).
- [51] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *CHI*.
- [52] Kaivalya Rawal and Himabindu Lakkaraju. 2020. Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses. In *NeurIPS*.
- [53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD*.
- [54] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* (2019).
- [55] Ivan Sanchez, Tim Rocktaschel, Sebastian Riedel, and Sameer Singh. 2015. Towards extracting faithful and descriptive representations of latent variable models. *AAAI Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches* (2015).
- [56] Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2021. GLocalX-From Local to Global Explanations of Black Box AI Models. *Artificial Intelligence* 294 (2021).

- [57] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *ICML*.
- [58] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*.
- [59] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.
- [60] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *AIES*.
- [61] Il'ya Meerovich Sobol'. 1990. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie* 2, 1 (1990).
- [62] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 41, 3 (2014).
- [63] Sarah Tan. 2018. Interpretable Approaches to Detect Bias in Black-Box Models. In *AIES Doctoral Consortium*.
- [64] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In *AIES*.
- [65] Sarah Tan, Matvey Soloviev, Giles Hooker, and Martin T. Wells. 2020. Tree space prototypes: Another look at making tree ensembles interpretable. In *FODS*.
- [66] Michael Tsang, Dehua Cheng, and Yan Liu. 2018. Detecting statistical interactions from neural network weights. In *ICLR*.
- [67] Ilse van der Linden, Hinda Haned, and Evangelos Kanoulas. 2019. Global aggregations of local explanations for black box models. *SIGIR Fairness, Accountability, Confidentiality, Transparency, and Safety Workshop* (2019).
- [68] Brian Williamson and Jean Feng. 2020. Efficient nonparametric statistical inference on population feature importance using Shapley values. In *ICML*.
- [69] Simon N. Wood. 2006. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.
- [70] Simon N. Wood. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B* 73, 1 (2011).
- [71] Tom Yan and Ariel D Procaccia. 2021. If you like shapley then you'll love the core. In *AAAI*.
- [72] Qingyuan Zhao and Trevor Hastie. 2021. Causal interpretations of black-box models. *Journal of Business & Economic Statistics* 39, 1 (2021).