Learning Global Additive Explanations of Black-Box Models

Sarah Tan*
ht395@cornell.edu
Cornell University

Rich Caruana
rcaruana@microsoft.com
Microsoft Research

Giles Hooker
gjh27@cornell.edu
Cornell University

Paul Koch
paulkoch@microsoft.com
Microsoft Research

Albert Gordo
albert.gordo.s@gmail.com

ABSTRACT

Interpretability has largely focused on local explanations, i.e. explaining why a model made a particular prediction for a sample. These explanations are appealing due to their simplicity and local fidelity, but do not provide information about the overall behavior of the model. We propose to use distillation to learn global additive explanations that describe the relationship between input features and model predictions. Unlike other global explanation methods, distillation allows us to learn explanations in a discriminative manner, minimizing the fidelity error between the black-box model and the explanation while preserving the explanation's interpretability. These global additive explanations take the form of feature shapes, which are more expressive than feature attributions. Through careful experimentation, including a user study on expert users, we show qualitatively and quantitatively that learned global additive explanations are able to describe model behavior and yield insights about black-box models.

1 INTRODUCTION

Recent research in interpretability has focused on developing *local* explanations: given an existing model and a sample, explain why the model made a particular prediction for that sample [38]. The accuracy and quality of these explanations have rapidly improved, and they are becoming important tools in interpretability. However, the human cost of examining multiple local explanations can be prohibitive, and it is unclear whether multiple local explanations can be aggregated without contradicting each other [1, 39].

In this paper, we are interested in *global* explanations: given an existing model, describe the overall behavior of the model. We operationalize this goal as describing the relationship between model inputs (features) and outputs (predictions), which is fundamental for several key tasks, such as understanding which features are important or debugging unexpected relationships learned by the model. As this task is most meaningful when each feature has semantic meaning [13], we focus on tabular data in this paper.

Given the prediction function of a black-box model, $F(\mathbf{x})$ and samples $\mathbf{x}$ consisting of features $x_i, \ldots, x_p$, we propose to use model distillation techniques [9, 22] to learn post-hoc global additive explanations of the form

$$\hat{F}(\mathbf{x}) = h_0 + \sum_i h_i(x_i) + \sum_{i \neq j} h_{ij}(x_i, x_j) + \sum_{i \neq j} \sum_{j \neq k} h_{ijk}(x_i, x_j, x_k) + \cdots \tag{1}$$

to approximate the model's prediction function $F(\mathbf{x})$. Figure 1 illustrates the approach. To summarize, the black-box model is treated as a teacher and distilled into a student (an additive model) that can be visualized as a set of feature shapes $\{h_i\}$, $\{h_{ij}\}$, $\{h_{ijk}\}$. Individual feature shapes can then be examined to determine the relationship between that feature and model predictions, the goal of our global explanations.

Feature shapes are not a new concept. Partial dependence [16], a classic post-hoc explanation method, and additive models learned directly on data [20] are also visualized in the form of feature shapes. The advantage of our approach over other additive explanations such as partial dependence (that is not learned using model distillation) is that distillation explicitly minimizes the error between the black-box model and the explanation, hence increasing the fidelity of the learned explanation. Distilling or approximating a black-box model by a interpretable model to serve as a global explanation is also not a new concept [6, 12, 18, 19, 28, 40]. We show, with experiments on expert users (machine learning model builders) that additive explanations have interpretability advantages over decision trees for certain model understanding tasks, and hence can be a viable explanation alternative.

When learning and evaluating post-hoc explanations, some questions naturally arise: how can we tell if the explanations are telling us something real about the black-box? Our paper answers this question by designing ground-truth explanations that we then show that our approach recovers.

The main contributions of this paper are:

- We learn global additive explanations for complex, non-linear models such as neural nets by coupling model distillation with powerful additive models to learn feature shapes that directly describe the relationship between features and predictions.
- We perform a quantitative comparison of our learned explanations to other *global* explanation methods. We measure fidelity to the black-box model as a function of the complexity of the explanation model, accuracy of the explanation on independent test data, and interpretability of the explanation.

*This work was performed during an internship at Microsoft Research.

Under review, do not distribute,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

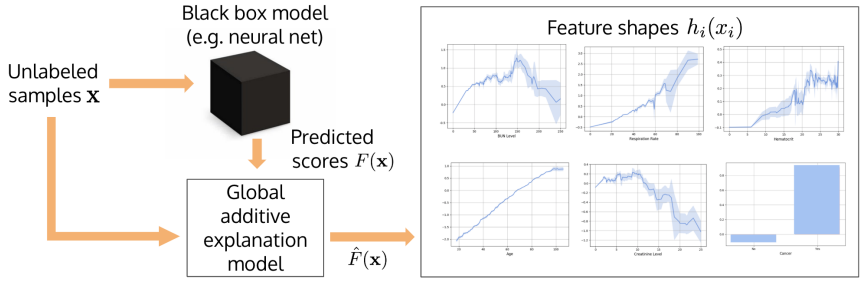

Figure 1: Given a black-box model and unlabeled samples (new unlabeled data or training data with labels discarded), our approach uses model distillation to learn feature shapes that describe the relationship between features and model predictions.

The results suggest that overall, additive explanations have higher fidelity with less complexity and have interpretability advantages over decision trees and linear models for certain model understanding tasks.

- Through a user study with expert users, we quantitatively measure how interpetable different global models are and how much they help users understand black-box models.

2 RELATED WORK

Global explanations. Neural nets and other black-box models have been approximated by interpretable models such as trees, rule lists [3, 31], decision sets [27], etc. either via model distillation [12, 18] or model extraction [6, 19, 28, 40]. All of these approximated classifiers; there has been less work approximating regression models. Craven and Shavlit [12] distilled a neural net into a decision tree and evaluated the explanation in terms of fidelity, accuracy, and complexity. Frosst and Hinton [18] also distilled a neural net into a soft decision tree. Neither evaluated interpretability of their explanations. In recent work, Lakkaraju et al. [27] extracted decision set explanations customized to the features the user is interested in.

Additive explanations. Several additive explanations, not learned via distillation, have been proposed [16, 24, 34, 42]. A common theme of these methods is that they *decompose* F into $\hat{F}$ using numerical or computational methods [24, 42] (e.g. matrix inversion, quasi Monte Carlo) which can be prohibitively expensive with large n or p, or permute features and repeatedly query the black-box model with the new data [16, 34], again a computationally expensive operation we avoid by learning explanations using distillation.

Feature attribution metrics. Several metrics have been proposed for feature importance for black-box models. These include permutation-based metrics [7], gradients/saliency (see [35] or [2] for a review), or metrics based on variance decompositions [25] or game theory [13, 34]. These metrics provide relative rankings of features but, as they do not characterize the full relationship between features and predictions, cannot answer questions such as "when feature x_i increases by 1, how does the prediction change?", which our explanations are able to answer.

Evaluation of interpretability. There is no universal definition of interpretability [14]; 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) [36, 37]; the exception is work mentioned above by Lakkaraju et al. [29] and concurrent work by Bastani et al. [6]; like us, they evaluate interpretability of global explanations on expert users.

3 OUR APPROACH

Our goal is to learn an explanation $\hat{F}$ that (1) describes the relationship between input features $x_1, \ldots, x_p$ and the model's prediction function F; (2) approximates prediction function F well.

3.1 Learning Global Additive Explanations

Treating the black-box model as a teacher, we use model distillation techniques [5, 9, 22] to learn global additive explanations for the black-box model.

Black-box model: fully-connected neural nets. Our black-box models are fully-connected neural nets (FNNs) with ReLU nonlinearities (see Appendix for training procedure). Note that our approach is not limited to neural nets, but can also be applied to learn explanations for other black-box models such as gradient boosted trees, random forests, etc. The most accurate nets we trained were FNNs with 2-hidden layers and 512 hidden units per layer (2H-512,512); nets with three or more hidden layers had lower training loss, but did not generalize as well on our data sets. In some experiments we also used a restricted-capacity model with 1 hidden layer of 8 units (1H-8). We obtain the prediction function of the black-box model, F, by having the black-box model label a set of training data.

Referring back to Equation (1), **additive explanations** are determined by the choice of metric L between F and its approximation $\hat{F}$, degree d of highest order components (e.g. $d = 3$ in Equation (1)), and type of base learner h. Learning $\hat{F}$ using model distillation is equivalent to choosing metric L that minimizing $||F - \hat{F}||_L$, the empirical risk between the prediction function F and our global additive explanation $\hat{F}$ on the training data.

Our choice of two flexible, nonparametric base learners for h – splines [44] and bagged trees – gives us two global additive explanation models $\hat{F}$: **Student Bagged Additive Boosted Trees (SAT)** and **Student Additive Splines (SAS)**. In addition, we include not just main components h_i but also higher order components h_{ij} and h_{ijk} to capture any interactions between features learned by the black-box model F and increase the fidelity of the explanation $\hat{F}$

to black-box model F. Throughout this paper, we call SAT with second-order components h_{ij} **SAT+pairs** and similarly for SAS. To train SAT, SAS, and SAT+pairs, we find optimal feature shapes $\{h_i\}$ and $\{h_{ij}\}$ that minimize mean square error between the black-box model F and the explanation $\hat{F}$, i.e.

$$L(h_0, h_1, \ldots, h_p) = \frac{1}{T}\sum_{t=1}^{T} \|F(x^t) - \hat{F}(x^t)\|_2^2$$
$$= \frac{1}{T}\sum_{t=1}^{T} \|F(x^t) - (h_0 + \sum_{i=1}^{p} h_i(x_i^t)) - \sum_{i\neq j} h_{ij}(x_i^t, x_j^t))\|_2^2, \quad (2)$$

where $F(x)$ is the black-box model's output (scores for regression tasks and logits for classification tasks), T is the number of training samples, x^t is the t-th training sample, and x_i^t is its i-th feature. The exact optimization details depend on the choice of h (see Appendix for training procedure).

3.2 Visualizing Global Additive Explanations

Our global additive explanations, SAT and SAS, can be visualized as **feature shapes** (Figure 1). These are plots with the x-axis being the domain of input feature x_i and the y-axis being the feature's contribution to the prediction $h_i(x_i)$. Feature shapes of SAT+pairs are heatmaps of x_i and x_j, with heatmap values being the two features' interaction contribution to the prediction $h_{ij}(x_i, x_j)$. As mentioned in Section 1, this way of representing the relationship between features and model predictions has precedence in interpretability, with additive models learned directly on data [20] and other additive explanations (not learned using model distillation) such as partial dependence [16] and Shapley additive explanations [34] also visualized in the form of feature shapes.

Given that the **visual complexity** of additive explanations is similar – one feature shape per feature – we compare our global additive explanations to partial dependence and Shapley additive explanations in terms of fidelity (Section 4.2). However, an interesting question arises in terms of how to fairly compare additive explanations and non-additive explanations such as distilled decision trees, sparse linear models, etc., with each explanation having different representations and hence different visual complexity. We do so with a comparison of fidelity, visual complexity, and interpretability in Sections 4.3 and 5.

4 EXPERIMENTAL RESULTS

The motivation and intended use case of global explanations described in Section 1 suggests the following criteria to evaluate our learned explanations:

(1) **Correctness**: do learned explanations look like ground-truth explanations, if available?
(2) **Fidelity**: are learned explanations faithful to the black-box model?
(3) **Complexity**: how does complexity affect the fidelity and interpretability of learned explanations?
(4) **Interpretability**: Can humans use the learned explanations to understand the overall behavior of the black-box model?

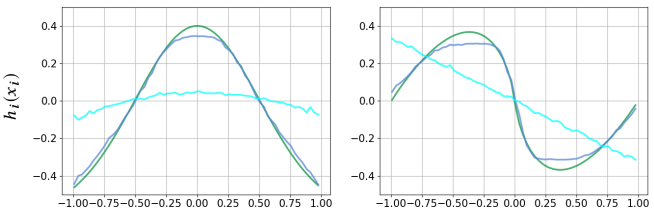

Figure 2: Comparison of ground truth, SAT of a 2H-512,512 black-box model, and SAT of a 1H-8 black-box model feature shapes for two representative features of F_1: x_4 (left) and x_6 (right).

Model	All	Agree	Disagree
2H-512,512	0.142	0.141	0.180
1H-8	0.483	0.407	0.489

Table 1: RMSE error of the 2H and 1H black-box models on all samples, compared to the error on samples sampled from regions where the explanation feature shapes "agree" or "disagree" with the ground truth shape.

4.1 Evaluating Correctness: Synthetic Data with Ground-Truth Explanations

In this experiment, we simulate ground-truth descriptions of feature-prediction relationships to see if our explanations can correctly recover them.

Setup. Inspired by [17], we designed an additive, highly non-linear function combining components from synthetic functions proposed by [17], [23] and [43]: $F_1(\mathbf{x}) = 3x_1 + x_2^3 - \pi^{x_3} + \exp(-2x_4^2) + \frac{1}{2+|x_5|} + x_6\log(|x_6|) + \sqrt{2|x_7|} + \max(0, x_7) + x_8^4 + 2\cos(\pi x_8)$. Like [43], we set the domain of all features to be Uniform[-1,1]. Like [17], we add noise features to our samples that have no effect on $F_1(x)$ via two noise features x_9 and x_{10}. We simulate 50,000 samples, and train two neural nets, 2H-512,512 and 1H-8, to predict F_1 from the ten features.

Performance of black-box model and explanations. The high-capacity 2H neural net obtained a test accuracy RMSE of 0.14, while the low-capacity neural net obtained test accuracy RMSE of 0.48, more than 3x larger, showing that function F_1 is not trivial. We trained a SAT global additive explanation[1] for each neural net. SAT explanations are faithful, with a fidelity RMSE of 0.14 to the 1H neural net, and a fidelity RMSE of 0.08 to the 2H neural net.

Does SAT explain the black-box model, or just the original data? A first question one may have when learning post-hoc explanations of black-box models is whether the learned explanation is describing relationships encoded in the black-box model or relationships in the original data. Figure 2 compares the feature shapes of our SAT explanation to function F_1's analytic ground-truth feature shapes for two features, x_4 and x_6, of F_1 (the behavior for other features is similar). We make two observations. First,

[1]We also experimented with SAS and obtained very similar results. For brevity and simplicity, in this section, we report only the results obtained by SAT.

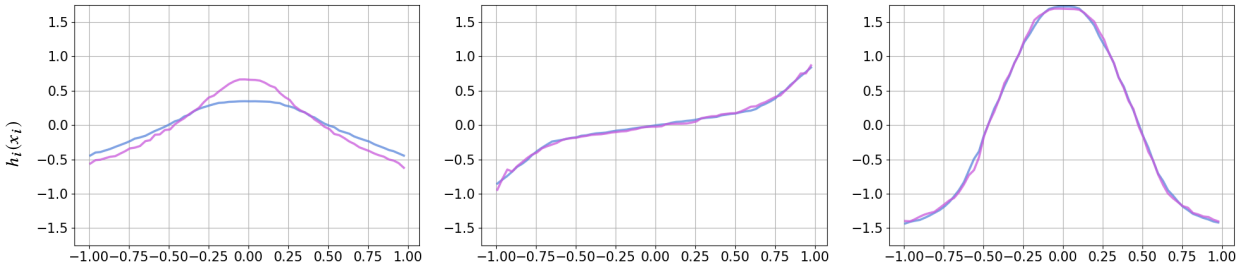

Figure 3: Comparison of SAT of a 2H-512,512 black-box model on F_1, and SAT of a 2H-512,512 black-box model on F_2 feature shapes on 3 features: x_4 (left), x_2 (center), and x_8 (right). Both x_2 and x_4 participate in interactions in F_2, while x_8 does not.

SAT's shapes for the 2H black-box model largely match the ground-truth shapes. Second, SAT's shapes for the 1H black-box model are notably different than the shapes for the 2H model, and are also less similar to the ground truth shapes. The differences in the SAT shapes for the 1H and 2H black-box models, combined with the accuracy of the black-box models and the similarity of the explanations to the ground truth, clearly indicate that the explanations explain the black-box models and not the underlying data.

Does SAT's feature shapes match the real behavior of the black-box model? We address this question two-ways. First, we directly measure the fidelity of SAT explanations, and compare it with the accuracy of the black-box models: the 2H black-box model has an accuracy of 0.14 RMSE, and its SAT explanation has a fidelity of 0.08 RMSE; the 1H black-box model has an accuracy of 0.48 RMSE, and its SAT explanation has a fidelity of 0.14. The fidelity of the explanations is significantly better than the black box models' accuracies, indicating that the explanations are faithful to the black-box models. Second, we measure the black box model's accuracy on samples belonging to regions where the explanations and the ground truth agree or disagree – for example, for the 2H model, $x_4 \approx 0$ and $|x_6| \approx 0.3$ define a region where the predicted feature shapes and the ground truth feature shapes disagree. If the SAT feature shapes accurately represent the black-box model, then the black-box model accuracy should be better on points sampled from areas of agreement than on points sampled from areas of disagreement. We confirm this behavior in Table 1: points sampled on the disagreement regions have lower accuracy than points sampled from the agreement regions[2].

How do interactions between features affect the feature shapes? We design an augmented version of F_1, $F_2(\mathbf{x}) = F_1(\mathbf{x}) + x_1x_2 + |x_3|^{2|x_4|} + \sec(x_3x_5x_6)$, which introduces interactions for features x_1 to x_6, to investigate how interactions in the black-box model's predictions are expressed by feature shapes. We simulate 50,000 samples, and train a new 2H-512,512 neural net to predict F_2 from the ten features. This function is much harder to learn (the 2H model obtained an RMSE of 0.21, compared to 0.14 of F_1) and also harder for explanation models (fidelity RMSEs of 0.35, compared to 0.08 RMSE of F_1).

[2]Note that, for 2H, the explanation matches the ground truth on most points, hence the accuracy of All is similar to the accuracy of Agree. For 1H, the explanation does not match the ground truth on most points, hence the accuracy of All is similar to the Accuracy of Disagree.

					Performance	
Data	n	p	**Type**		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 neural nets. For RMSE, lower is better. For AUC, higher is better.

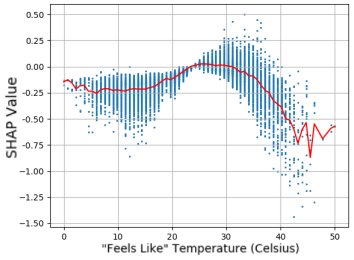

Figure 4: From local Shapley explanations to gSHAP. Blue points are local Shapley explanations; red line is a global gSHAP feature shape.

Figure 3 displays the feature shapes of the SAT explanations from F_2 (in purple) for two features with interactions (x_4, x_2) and a feature without interactions (x_8), and compares them with the shapes from F_1 (in blue), already discussed in Figure 2. We first note how, for x_8 (right), the shapes from F_1 and F_2 match almost perfectly: the explanation model was not confused by the other interactions and was able to accurately match the shape of x_8. For x_4 (left), the part of the interactions that can be approximated additively by h_i has "leaked" into the h_i feature shape, slightly changing its shape as expected.

An interesting case is x_2, where, despite interacting with x_1, its feature shape has not changed and matches the feature shape from F_1. This is less surprising if we recall that feature shapes describe the *expected importance* of the feature, learned in a data-driven fashion. The interaction term is x_1x_2, which, for $x_1 \sim$ Uniform[-1,1], has an expected value of zero, and therefore does not affect the feature shape. Similarly, for x_4, the expected value of the interaction $|x_3|^{2|x_4|}$ when $x_3 \sim$ Uniform[-1,1] is $1/(2|x_4| + 1)$, an upward pointing cusp, which leads to the change noticed in Figure 3 (left).

Accuracy	Global Explanation	Bikeshare RMSE	Loan score RMSE	Magic AUC	Pneumonia AUC	FICO AUC
Ours	SAT	0.98 ± 0.00	2.35 ± 0.01	90.75 ± 0.06	82.24 ± 0.05	79.42 ± 0.04
	SAT+pairs	0.60 ± 0.00	2.13 ± 0.01	90.75 ± 0.06	82.23 ± 0.06	79.44 ± 0.04
	SAS	0.98 ± 0.00	2.34 ± 0.00	90.58 ± 0.02	82.12 ± 0.04	79.51 ± 0.02
Other additive methods	gGRAD	1.25 ± 0.00	6.04 ± 0.01	80.95 ± 0.13	81.88 ± 0.05	79.28 ± 0.02
	gSHAP	1.02 ± 0.00	5.10 ± 0.01	88.98 ± 0.05	82.31 ± 0.03	79.36 ± 0.01
	PD	1.00 ± 0.00	4.31 ± 0.00	82.78 ± 0.00	82.15 ± 0.00	79.47 ± 0.00
Other interpretable methods	Decision Tree	0.60 ± 0.01	2.66 ± 0.02	91.44 ± 0.29	79.38 ± 0.38	78.19 ± 0.03
	Sparse Linear	1.39 ± 0.00	3.45 ± 0.00	86.91 ± 0.01	82.06 ± 0.02	79.16 ± 0.01
Fidelity	Global Explanation	Bikeshare RMSE	Loan score RMSE	Magic RMSE	Pneumonia RMSE	FICO RMSE
Ours	SAT	0.92 ± 0.00	1.74 ± 0.01	1.78 ± 0.00	0.35 ± 0.00	0.15 ± 0.00
	SAT+pairs	0.50 ± 0.00	1.47 ± 0.00	1.75 ± 0.00	0.30 ± 0.00	0.11 ± 0.00
	SAS	0.92 ± 0.00	1.71 ± 0.00	1.75 ± 0.00	0.35 ± 0.00	0.14 ± 0.00
Other additive methods	gGRAD	1.20 ± 0.00	5.93 ± 0.01	2.93 ± 0.01	0.43 ± 0.00	0.16 ± 0.00
	gSHAP	0.96 ± 0.00	4.83 ± 0.00	2.15 ± 0.00	0.46 ± 0.00	0.16 ± 0.00
	PD	0.94 ± 0.00	3.85 ± 0.00	3.17 ± 0.00	0.47 ± 0.00	0.16 ± 0.00
Other interpretable methods	Decision Tree	0.47 ± 0.01	2.12 ± 0.02	1.33 ± 0.03	0.75 ± 0.01	0.44 ± 0.01
	Sparse Linear	1.35 ± 0.00	2.87 ± 0.01	2.22 ± 0.00	0.49 ± 0.00	0.18 ± 0.00

Table 3: Accuracy and fidelity of global additive explanations for 2H neural nets. Accuracy is RMSE for regression tasks and AUROC for classification tasks; fidelity is always RMSE between the explanation model's predictions and the black-box model's scores or logits (see Equation (2)).

4.2 Evaluating Fidelity and Accuracy: Comparing Explanations on Real Data

In this section, we quantitatively compare our global additive explanations to other global 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 [30], the 2018 FICO Explainable ML Challenge's credit data set [15], and the pneumonia data set analyzed by [11]. We train a 2H-512,512 neural net that we will use as the main black-box model in this section (see Appendix for training procedure). Table 2 presents the accuracy of the 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.

Metrics. Lundberg and Lee [34] suggested viewing an explanation of a model's prediction as a model itself. With this perspective, we quantitatively evaluate explanation models as if they were models. Specifically, we evaluate not just fidelity (how well the explanation matches the black-box model's predictions) but also accuracy (how well the explanation predicts the original label). Note that [34] and [38] evaluated local fidelity (called local accuracy by [34]), but not accuracy. A similar evaluation of global accuracy was performed by [26] who used their explanations (prototypes) to classify test data. We use the feature shapes of additive explanations and distilled interpretable models to predict on independent test data.

Baselines. We compare to two types of baselines (see Appendix for training procedure): (1) Additive explanations obtained by querying the black-box model (i.e. without distillation): partial dependence (PD) [16], Shapley additive explanations [34] and gradient-based explanations [41]; (2) Interpretable models learned by distilling the black-box model: trees and sparse linear models.

Both Shapley additive explanations and gradient-based explanations are local explanations that we adapt to a global setting by averaging the local explanations at each unique feature value. For example, the global attribution for feature "Temperature" at value 10 is the average of local attribution "Temperature" for all training samples with "Temperature=10". This is the red line passing through the points in Figure 4. Applying this procedure to Shapley and gradient-based local attributions, we obtain global attributions **gGRAD** and **gSHAP** that we can now plot as feature shapes.

Table 3 presents the fidelity and accuracy results for SAT and SAS compared to the two types of baselines: (1) other additive explanations; (2) other distilled interpretable models. We also include an augmented version of SAT that includes pairwise interactions, denoted by SAT+pairs.

We draw several conclusions. First, SAT and SAS yield similar results in all cases, both in terms of accuracy and fidelity, indicating that the particular choice of the base learner did not matter for these data sets. Capturing pairwise interactions (SAT+pairs) leads to improvements in some datasets (particularly Bikeshare and Loan, the two regression tasks), while in the remaining datasets the changes are not as remarkable. This suggests that the individual feature shapes already provide a faithful interpretation of the model.

Compared to other additive explanations such as gSHAP and PD, SAT and SAS generally obtain better accuracy and fidelity. This is not surprising since SAT and SAS were trained specifically to mimic the black-box model. In particular, SAT and SAS are superior to PD in all tasks and metrics. Compared to other interpretable

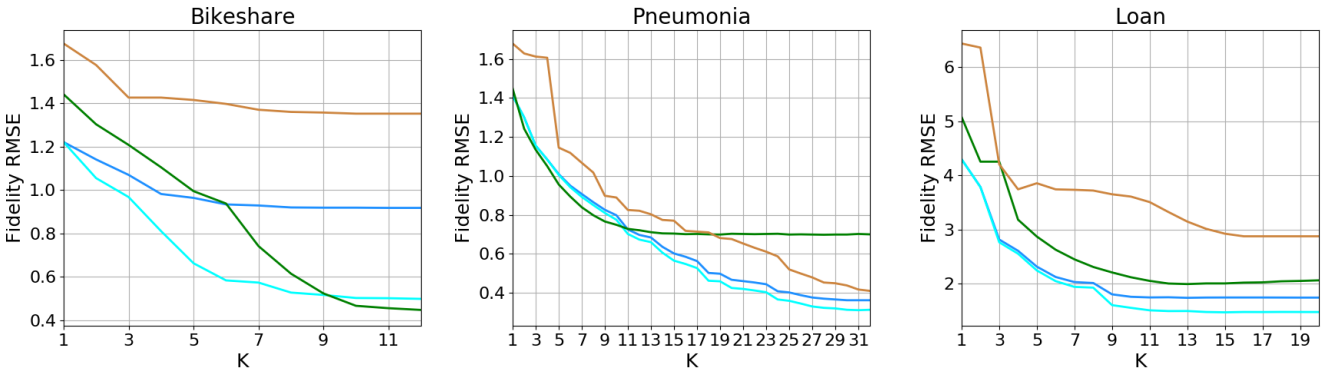

Figure 5: Fidelity (RMSE) of different distilled interpretable models on Bikeshare (left), Pneumonia (center) and Loan (right) datasets as a function of model "complexity" K (number of features for SAT and SPARSE, tree depth for DR). The lower the fidelity RMSE, the more faithful the interpretable model to the black-box model. Legend: SAT, SAT+Pairs, DT, SPARSE

methods, SAS and SAT also obtain better results. Despite not capturing interactions, SAS and SAT are non-linear models, and hence able to model nonlinear relationships that sparse linear models cannot. Decision trees are locally adaptive smoothers [8] better able to adapt to sudden changes in input-output relationships, but that also gives them more capacity to overfit. They excel on some datasets (e.g. Bikeshare), but are not as accurate on other datasets (e.g. Pneumonia or FICO).

4.3 Evaluating Fidelity as a Function of Explanation Complexity

In the previous section we compared the fidelity and accuracy of SAT and SAS to other additive explanations such as gGRAD, gSHAP, and PD. Because all these methods are additive they can be visualized in feature shapes. Models such as trees and rules, however, may be interpretable but are not additive. In this section we compare the *fidelity* of SAT explanations to sparse linear models and trees of varying complexity, showing that **the most faithful models with low complexity may be different from the most faithful models with high complexity.** In Section 5 we then compare the *interpretability* of SAT to trees and linear models via a user study, tying the complexity of the models with their actual interpretability.

Figure 5 presents the fidelity[3] of SAT and SAT+pairs compared to two other interpretable distilled models, decision trees (DT) and sparse L1-regularized linear models (SPARSE), on three of the test problems: Bikeshare, Pneumonia and Loan. The trees and linear models are trained using scikit-learn[4].

We present results as a function of a model-specific parameter K that controls the complexity of the model. For SPARSE, K represents the number of features included in the model, controlled indirectly through the LASSO regularization parameter α. For DT, K is the depth of the tree. We allow a tree of depth K access to all features. Because of this, a tree of depth K might use fewer than K features (continuous or multi-valued features might be split more than once on some branches), exactly K features (e.g., if all features are Boolean), or more than K features (by splitting different features on different branches — the most common case). For SAT and SAT+pairs, K is the number of features included in the additive model. For SAT this is also the number of shape plots. For SAT+pairs, which models pairwise interactions, the model will also include shape plots that represent stronger pairwise interactions found between the K features in the model. Note that trees of depth K can represent K-way interactions, and that the model complexity of trees falls between K and 2^K because a binary tree of depth K has 2^K leaves (2^K rules), but the complexity is somewhat less than 2^K because there is overlap in the rules resulting from the tree structure.

Overall, SPARSE has the worst fidelity. On Bikeshare and Loan, SPARSE is dominated by all other methods. On Pneumonia it is inferior to SAT and SAT+pairs for all values of K, but has better or worse fidelity than trees depending on K. Even though linear models may be interpretable, they often do not have the complexity necessary to accurately represent most black-box models. Note that two explanation methods that use sparse linear models [38] and rules [39] use them as local (not global) explanations, and only for classification (not regression).

Trees perform well given enough features and depth. On Bikeshare, trees outperform SAT by depth 7, and outperform SAT+pairs by a small amount for depth 10 and greater, although at that point one has to consider up to $2^{10} = 1024$ different paths. We suspect the deep tree is able to benefit from higher order interactions, whereas SAT+pairs is restricted to pairwise interactions to maintain intelligibility. However, the user study in Section 5 suggests that trees of this depth are no longer intelligible.

Overall, the best model is SAT+pairs. On Bikeshare, where interactions are important, SAT+pairs performs much better than SAT (which uses the same features but does not model interactions between features), and outperforms shallow trees of depth 8 or less. On Pneumonia and Loan, both SAT and SAT+pairs outperform

[3]The accuracy plots present very similar patterns.

[4]We also tried to compare to rule lists. However, state-of-the-art rule lists [3, 31] do not support regression, which is needed for distillation. We considered a slightly older subgroup discovery algorithm [4] that supports regression but does not generate disjoint rules, but we only achieved reasonable results on the Bikeshare dataset, hence we preferred not to report the rules results. We however use these rules for our user study on Bikeshare in Section 5.

SPARSE and trees of any depth. SAT+pairs consistently outperforms SAT on all three problems, by wide margin on Bikeshare, and small margins on Pneumonia and Loan.

In summary, our global additive explanations (SAT and SAT+pairs) have the highest overall fidelity to the black-box models they are trained to explain, even at low values of K. Trees sometimes exhibit high fidelity when given adequate depth, but the results from the user study in the next section suggest that depth greater than 5 or 6 hinders their intelligibility.

5 EVALUATING INTERPRETABILITY WITH EXPERT USERS

We now describe the results from a user study to see if SAT additive explanations can be understood and used by humans, comparing them to other interpretable models (DT, SPARSE, RULES) distilled from the 2H-512,512 neural net. We denote the complexity of the models by model-K. For example, a tree of depth 4 would be denoted as DT-4, while a group of 5 rules would be denoted as RULES-5. Table 4 presents quantitative results from the user study.

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 (see Section B) 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.

In the first stage, 24 of 50 subjects were randomly assigned to see output from DT-4 or SAT-5[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 models (in terms of accuracy and fidelity): sparse linear models (SPARSE-2) and subgroup-rules (RULES-5).

Can humans understand and use feature shapes? From the absolute magnitude of the SAT feature shapes 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 A2). 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.

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 shape in Figure A1).

How do tree depth and number of feature shapes affect human performance? We also experimented with smaller models, SAT-2 and DT-2, that used only the two most important features, Hour and Temperature. As the models 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 fidelity and 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 model, but increases it by 166% for the DT model, 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 interpretable models: 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,

[5] We considered DT and SAT first because they are the most accurate and faithful explanations. We used 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 (RMSE: SAT 0.98, DT-6 1, DT-4 1.16). DT-4 used five features: Hour, Temperature, Year, Working Day, Season (Figure A2), hence we select the corresponding five feature shapes to display for SAT-5 (Figure A1).

	First stage (n=24)		Second stage (n=14)		Third stage (n=12)	
Task	**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

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.

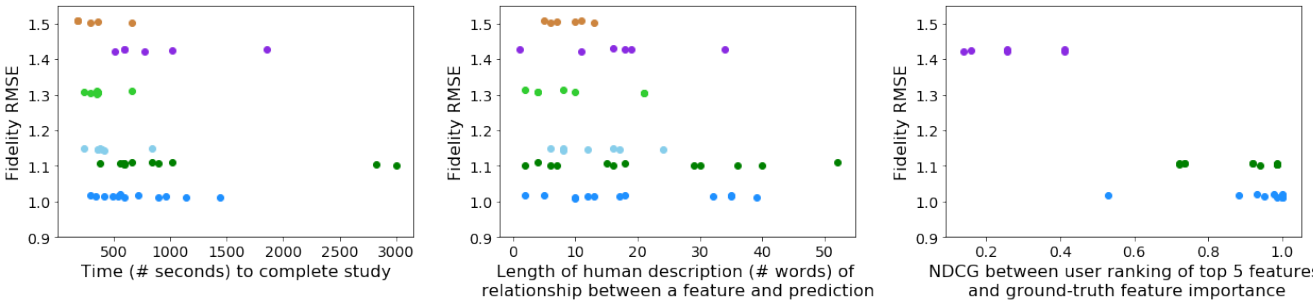

Figure 6: Time needed to finish the study (left), length of the description (center), and the NDCG of the ranked features (right), compared to the Fidelity RMSE, for the individual users and methods. Legend: SAT-5, DT-4, SAT-2, DT-2, RULES-5, SPARSE-2

this was the fastest explanation to interpret. However, the model 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 6 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 [14]. On the y-axis is fidelity (RMSE).

The plots show that there is a trade-off between fidelity and interpretbility (as measured by time to complete, description length, and NDCG of feature rankings), but not all methods behave similarly. In general, the SPARSE-2 model 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.

6 CONCLUSIONS

We presented a method for "opening up" complex models such as neural nets trained on tabular data. The method, based on distillation with high-accuracy additive models, has clear advantages over other global explanations that learn additive explanations without distillation, and non-additive explanations such as trees that do use distillation. The method will work with any black-box classification or regression model including random forests and boosted trees, but is not designed for models such as CNNs trained on raw inputs such as images where providing a global explanation in terms of input pixels is not meaningful. Different kinds of explanations are useful for different purposes, and global additive models do not aim to replace local explanations. The results of our experiments and a user study on expert users (machine learning model builders) suggest that distillation into high-performance additive models provides explanations that have a strong combination of fidelity, low-complexity, and interpretability.

REFERENCES

[1] David Alvarez-Melis and Tommi S Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. In *NIPS*.

[2] Marco Ancona, Enea Ceolini, Cengiz ÃŰztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *ICLR*.

[3] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. 2017. Learning certifiably optimal rule lists. In *KDD*.

[4] M. Atzmueller and F Lemmerich. 2012. VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. http://www.vikamine.org/

[5] Jimmy Ba and Rich Caruana. 2014. Do Deep Nets Really Need to be Deep?. In *NIPS*.

[6] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2019. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504* (2019).

[7] Leo Breiman. 2001. Random forests. *Machine Learning* (2001).

[8] L Breiman, JH Friedman, RA Olshen, and CJ Stone. 1984. *Classification and regression trees (CART)*. Wadsworth International Group.

[9] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *KDD*.

[10] Peter Buhlmann and Bin Yu. 2003. Boosting with the L2 loss: regression and classification. *J. Amer. Statist. Assoc.* (2003).

[11] 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*.

[12] Mark W. Craven and Jude W. Shavlik. 1995. Extracting Tree-structured Representations of Trained Networks. In *NIPS*.

[13] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy*.

[14] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[15] FICO. 2018. Explainable Machine Learning Challenge. (2018). https://community.fico.com/s/explainable-machine-learning-challenge

[16] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* (2001).

[17] Jerome H Friedman and Bogdan E Popescu. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics* (2008).

[18] Nicholas Frosst and Geoffrey Hinton. 2017. Distilling a Neural Network Into a Soft Decision Tree. *arXiv preprint arXiv:1711.09784* (2017).

[19] LiMin Fu. 1994. Rule generation from neural networks. *IEEE Transactions on Systems, Man, and Cybernetics* (1994).

[20] Trevor Hastie and Rob Tibshirani. 1990. *Generalized Additive Models*. Chapman and Hall/CRC.

[21] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer.

[22] Geoff Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *NIPS Deep Learning Workshop* (2015).

[23] Giles Hooker. 2004. Discovering additive structure in black box functions. In *KDD*.

[24] Giles Hooker. 2007. Generalized Functional ANOVA Diagnostics for High Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics* (2007).

[25] Bertrand Iooss and Paul Lemaitre. 2015. A review on global sensitivity analysis methods. In *Uncertainty Management in Simulation-Optimization of Complex Systems*.

[26] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*.

[27] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *KDD*.

[28] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In *AIES*.

[29] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In *AIES*.

[30] LendingClub. 2011. Lending Club Loan Data. (2011). https://www.lendingclub.com/info/download-data.action

[31] 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* (2015).

[32] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *KDD*.

[33] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *KDD*.

[34] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *NIPS*.

[35] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2017. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* (2017).

[36] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint arXiv:1802.00682* (2018).

[37] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD*.

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI*.

[40] Ivan Sanchez, Tim Rocktaschel, Sebastian Riedel, and Sameer Singh. 2015. Towards extracting faithful and descriptive representations of latent variable models. *AAAI Spring Syposium on Knowledge Representation and Reasoning (KRR): Integrating Symbolic and Neural Approaches* (2015).

[41] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *ICML*.

[42] Ilya M Sobol. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation* (2001).

[43] Michael Tsang, Dehua Cheng, and Yan Liu. 2018. Detecting statistical interactions from neural network weights. In *ICLR*.

[44] Simon N Wood. 2006. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

[45] Simon N Wood. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* (2011).

APPENDIX

A Training procedure and implementation details

A.1 Black-box model: neural nets. The neural net training was done using PyTorch. We use the Adam optimizer (default beta parameters), Xavier initialization, 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, etc) based on average validation performance on multiple train-validation splits and random initializations.

A.2 Student additive explanations. For student additive explanations with tree base learners (SAT), we use cyclic gradient boosting [10, 32] which learns the feature shapes in a cyclic manner. As trees are high-variance, low-bias learners [21], when used as base learners in additive models, it is standard to bag multiple trees [11, 32, 33]. We follow that approach here. For student additive explanations with spline base learners (SAS), we use cubic regression splines trained using penalized maximum likelihood in R's `mgcv` library [45] and cross-validate the splines' smoothing parameters.

A.3 Baselines. **Partial dependence** [16] (PD) is a classic global explanation method that estimates how predictions change as feature x_j varies over its domain: $PD(x_j = z) = \frac{1}{T}\sum_{t=1}^{T} F((x_1^t, \ldots, x_j^t = z, \ldots, x_p^t)$ where the neural net is queried with new data samples generated by setting the value of their x_j feature to z, a value in the domain of x_j. Plotting $PD(x_j = z)$ by z returns a feature shape. We implement our own version of partial dependence by repeatedly setting x_j^t for all points to a, a value in the domain of x_j, and then querying the neural net with these new data samples.

Gradient-based explanations involves constructing the additive function G through the Taylor decomposition of F, defining $G(x) = F(0) + \sum_{i=1}^{p} \frac{\partial F(x)}{\partial x_i} x_i$, and defining the attribution of feature i of value x_i as $\frac{\partial F(x)}{\partial x_i} x_i$. This formulation is related to the "gradient*input" method (e.g. [41]) used to generate saliency maps for images.

Shapley additive explanations [34] is a state-of-the-art local explanation method that satisfies several desirable local explanation properties [34]. Given a sample and its prediction, Shapley additive explanations decompose the prediction additively between features using a game-theoretic approach. We use the python package by the authors of Shapley additive explanations.

Decision trees and Sparse linear models were learned using the scikit-learn Python package. **Subgroup rules** were learned using the Vikamine [4] package, as we needed to learn rules for regression problems and state-of-the-art rule lists [3, 31] do not support regression. However, our results with Vikamine were unsatisfying, and we only obtained reasonable results on the Bikeshare dataset.

B User Study Materials

All 50 user study subjects answered these questions:

(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.

The SAT-5 and DT-4 models shown to the users are found in Figures A1 and A2.

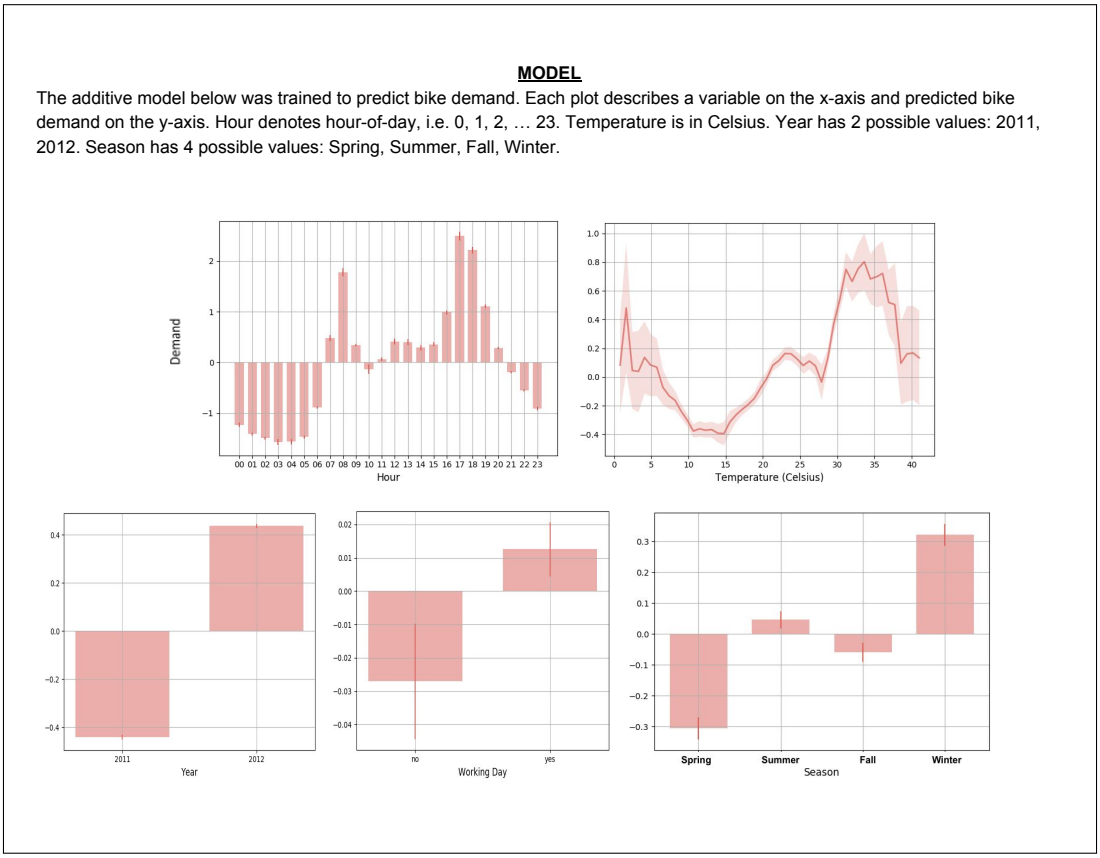

Figure A1: Model output shown to SAT-5 subjects.

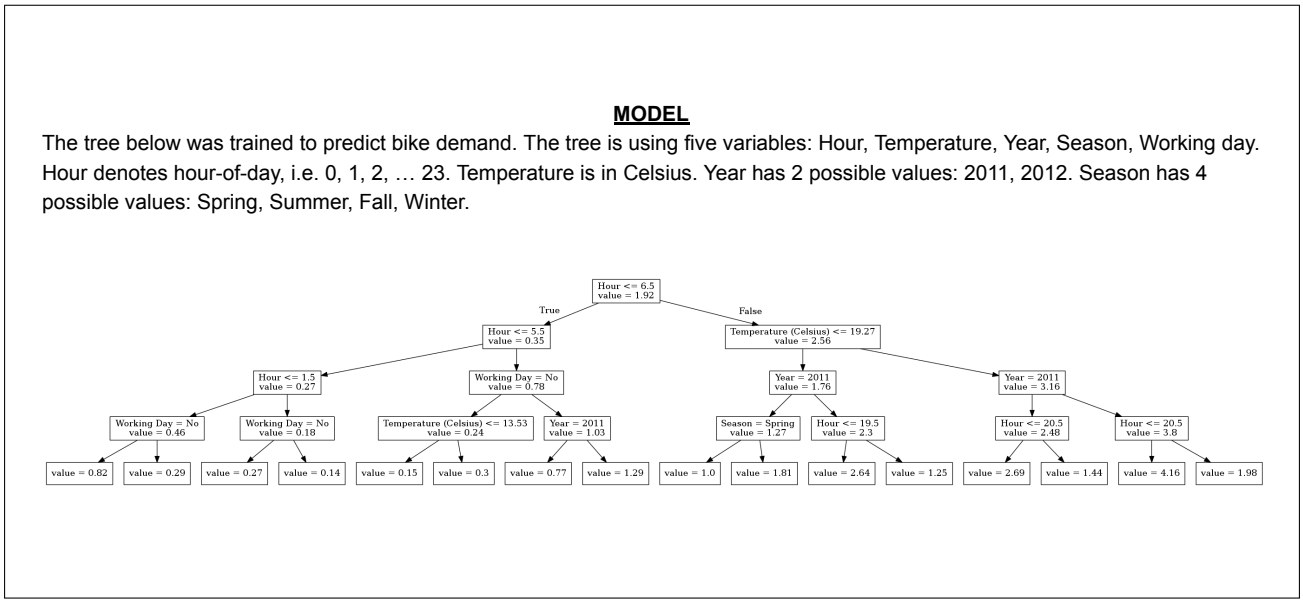

Figure A2: Model output shown to DT-4 subjects.