

Consistent Individualized Feature Attribution for Tree Ensembles

Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee

University of Washington
{slund1,erion,suinlee}@uw.edu

ABSTRACT

Interpreting predictions from tree ensemble methods such as gradient boosting machines and random forests is important, yet feature attribution for trees is often heuristic and not individualized for each prediction. Here we show that popular feature attribution methods are *inconsistent*, meaning they can lower a feature’s assigned importance when the true impact of that feature actually increases. This is a fundamental problem that casts doubt on any comparison between features. To address it we turn to recent applications of game theory and **develop fast exact tree solutions for SHAP** (SHapley Additive exPlanation) values, which are the unique consistent and locally accurate attribution values. We then extend SHAP values to **interaction effects and define SHAP interaction values**. We propose a rich visualization of individualized feature attributions that improves over classic attribution summaries and partial dependence plots, and a unique “supervised” clustering (clustering based on feature attributions). We demonstrate better agreement with human intuition through a user study, exponential improvements in run time, improved clustering performance, and better identification of influential features. An implementation of our algorithm has also been merged into XGBoost and LightGBM, see <http://github.com/slundberg/shap> for details.

1 INTRODUCTION

Understanding why a model made a prediction is important for trust, actionability, accountability, debugging, and many other tasks. To understand predictions from tree ensemble methods, such as gradient boosting machines or random forests, importance values are typically attributed to each input feature. These importance values can be computed either for a single prediction (individualized), or an entire dataset to explain a model’s overall behavior (global).

Concerningly, popular current feature attribution methods for tree ensembles are *inconsistent*. This means that when a model is changed such that a feature has a higher impact on the model’s output, current methods can actually lower the importance of that feature. Inconsistency strikes at the heart of what it means to be a good attribution method, because it prevents the meaningful comparison of attribution values across features. This is because inconsistency implies that a feature with a large attribution value might be less important than another feature with a smaller attribution (see Figure 1 and Section 2).

To address this problem we turn to the recently proposed SHAP (SHapley Additive exPlanation) values [16], which are based on a unification of ideas from game theory [27] and local explanations [21]. Here we show that by connecting tree ensemble feature attribution methods with the class of *additive feature attribution methods* [16] we can motivate SHAP values as the only possible consistent feature attribution method with several desirable properties.

SHAP values are theoretically optimal, but like other model agnostic feature attribution methods [2, 9, 21, 27], they can be challenging to compute. To solve this we derive an algorithm for tree ensembles that reduces the complexity of computing exact SHAP values from $O(TL^2M)$ to $O(TLD^2)$ where T is the number of trees, L is the maximum number of leaves in any tree, M is the number of features, and D is the maximum depth of any tree. This exponential reduction in complexity allows predictions from previously intractable models with thousands of trees and features to now be explained in a fraction of a second. Entire datasets can now be explained, which enables new alternatives to traditional partial dependence plots and feature importance plots [11], which we term *SHAP dependence plots* and *SHAP summary plots*, respectively.

Current attribution methods cannot directly represent interactions, but must divide the impact of an interaction among each feature. To directly capture pairwise interaction effects we propose *SHAP interaction values*; an extension of SHAP values based on the Shapley interaction index from game theory [12]. SHAP interaction values bring the benefits of guaranteed consistency to explanations of interaction effects for individual predictions.

In what follows we first discuss current tree feature attribution methods and their inconsistencies. We then introduce SHAP values as the only possible consistent and locally accurate attributions, present Tree SHAP as a high speed algorithm for estimating SHAP values of tree ensembles, then extend this to SHAP interaction values. We use user study data, computational performance, influential feature identification, and supervised clustering to compare with previous methods. Finally, we illustrate SHAP dependence plots and SHAP summary plots with XGBoost and NHANES I national health study data [18].

2 INCONSISTENCIES IN CURRENT FEATURE ATTRIBUTION METHODS

Tree ensemble implementations in popular packages such as XGBoost [6], scikit-learn [20], and the *gbm* R package [22] allow a user to compute a measure of feature importance. These values are meant to summarize a complicated ensemble model and provide insight into what features drive the model’s prediction.

Global feature importance values are calculated for an entire dataset (i.e., for all samples) in three primary ways:

- (1) Gain: A classic approach to feature importance introduced by Breiman et al. in 1984 [3] is based on gain. Gain is the total reduction of loss or impurity contributed by all splits for a given feature. Though its motivation is largely heuristic [11], gain is widely used as the basis for feature selection methods [5, 13, 25].
- (2) Split Count: A second common approach is simply to count how many times a feature is used to split [6]. Since feature

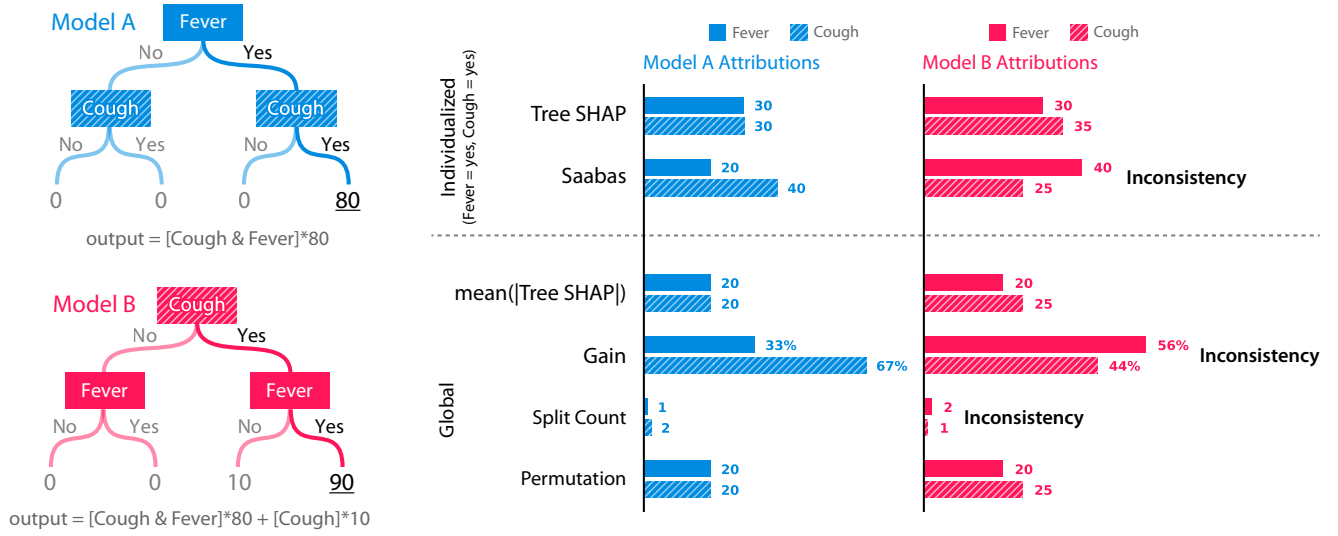


Figure 1: Two simple tree models that demonstrate inconsistencies in the Saabas, gain, and split count attribution methods: The Cough feature has a larger impact in Model B than Model A, but is attributed less importance in Model B. Similarly, the Cough feature has a larger impact than Fever in Model B, yet is attributed less importance. The individualized attributions explain a single prediction of the model (when both Cough and Fever are Yes) by allocating the difference between the expected value of the model’s output (20 for Model A, 25 for Model B) and the current output (80 for Model A, 90 for Model B). The global attributions represent the overall importance of a feature in the model. Without consistency it is impossible to reliably compare feature attribution values.

splits are chosen to be the most informative, this can represent a feature’s importance.

- (3) Permutation: A third common approach is to randomly permute the values of a feature in the test set and then observe the change in the model’s error. If a feature’s value is important then permuting it should create a large increase in the model’s error. **Different choices about the method of feature value permutation lead to variations of this basic approach** [1, 10, 14, 23, 26].

Individualized methods that compute feature importance values for a single prediction are **less established for trees**. While model agnostic individualized explanation methods [2, 9, 16, 21, 27] can be applied to trees [17], they are significantly slower than tree-specific methods and have sampling variability (see Section 5.3 for a computational comparison, or [16] for an overview). The **only current tree-specific individualized** explanation method we are aware of is **by Sabbas** [24]. The Saabas method is similar to the classic dataset-level gain method, but instead of measuring the reduction of loss, it measures the change in the model’s expected output. It proceeds by comparing the expected value of the model output at the root of the tree with the expected output of the subtree rooted at the child node followed by the decision path of the current input. The difference between these expectations is then attributed to the feature split on at the root node. By repeating this process recursively the method allocates the difference between the expected model output and the current output among the features on the decision path.

Unfortunately, the feature importance values from the gain, split count, and Saabas methods are all inconsistent. This means that a model can change such that it relies more on a given feature, yet the importance estimate assigned to that feature decreases. Of the methods we consider, only SHAP values and permutation-based methods are consistent. Figure 1 shows the result of applying all these methods to two simple regression trees.¹ For the global calculations we assume an equal number of dataset points fall in each leaf, and the label of those points is exactly equal to the prediction of the leaf. Model A represents a simple AND function, while Model B represents the same AND function but with an additional increase in the predicted value when Cough is “Yes”. Note that because Cough is now more important it gets split on first in Model B.

Individualized feature attribution is represented by Tree SHAP and Saabas for the input Fever=Yes and Cough=Yes. Both methods allocate the difference between the current model output and the expected model output among the input features (80 – 20 for Model A). But the SHAP values are guaranteed to reflect the importance of the feature (see Section 2.1), while the Saabas values can give erroneous results, such as a larger attribution to Fever than to Cough in Model B.

Global feature attribution is represented by four methods: the mean magnitude of the SHAP values, gain, split count, and feature permutation. Only the mean SHAP value magnitude and permutation correctly give Cough more importance than Fever in Model B.

¹For clarity we rounded small values in Figure 1. These small values are why the lower left splits in both models were not pruned during training.

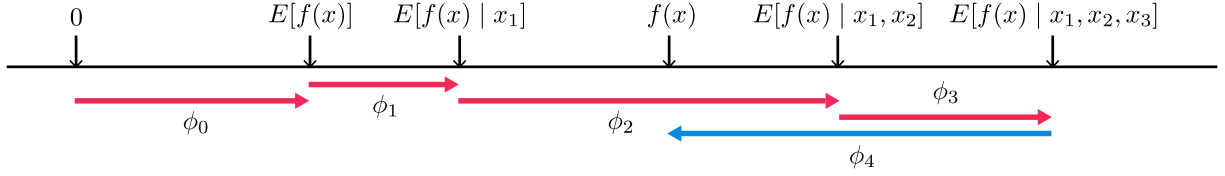


Figure 2: SHAP (SHapley Additive exPlanation) values explain the output of a function f as a sum of the effects ϕ_i of each feature being introduced into a conditional expectation. Importantly, for non-linear functions the order in which features are introduced matters. SHAP values result from averaging over all possible orderings. Proofs from game theory show this is the only possible consistent approach where $\sum_{i=0}^M \phi_i = f(x)$. In contrast, the only current individualized feature attribution method for trees satisfies the summation, but is inconsistent because it only considers a single ordering [24].

This means gain and split count are not reliable measures of global feature importance, which is important to note given their widespread use.

2.1 SHAP values as the only consistent and locally accurate individualized feature attributions

It was recently noted that many current methods for interpreting individual machine learning model predictions fall into the class of *additive feature attribution methods* [16]. This class covers methods that explain a model’s output as a sum of real values attributed to each input feature.

Definition 2.1. **Additive feature attribution methods** have an explanation model g that is a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i, \quad (1)$$

where $z' \in \{0, 1\}^M$, M is the number of input features, and $\phi_i \in \mathbb{R}$.

The z'_i variables typically represent a feature being observed ($z'_i = 1$) or unknown ($z'_i = 0$), and the ϕ_i ’s are the feature attribution values.

As previously described in Lundberg and Lee (2017), an important property of the class of additive feature attribution methods is that there is a single unique solution in this class with three desirable properties: *local accuracy*, *missingness*, and *consistency*. Local accuracy states that the sum of the feature attributions is equal to the output of the function we are seeking to explain. Missingness states that features that are already missing (such that $z'_i = 0$) are attributed no importance. Consistency states that changing a model so a feature has a larger impact on the model will never decrease the attribution assigned to that feature.

Note that in order to evaluate the effect missing features have on a model f , it is necessary to define a mapping h_x that maps between a binary pattern of missing features represented by z' and the original function input space. Given such a mapping we can evaluate $f(h_x(z'))$ and so calculate the effect of observing or not observing a feature (by setting $z'_i = 1$ or $z'_i = 0$).

To compute SHAP values we define $f_x(S) = f(h_x(z')) = E[f(x) | x_S]$ where S is the set of non-zero indexes in z' (Figure 2), and $E[f(x) | x_S]$ is the expected value of the function conditioned on

a subset S of the input features. SHAP values combine these conditional expectations with the classic Shapley values from game theory to attribute ϕ_i values to each feature:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)], \quad (2)$$

where N is the set of all input features.

As shown in Lundberg and Lee (2017), the above method is the only possible consistent, locally accurate method that obeys the missingness property and uses conditional dependence to measure missingness [16]. This is strong motivation to use SHAP values for tree ensemble feature attribution, particularly since the only previous individualized feature attribution method for trees, the Saabas method, satisfies both local accuracy and missingness using conditional dependence, but fails to satisfy consistency. This means that SHAP values provide a strict theoretical improvement by eliminating significant consistency problems (Figure 1).

3 TREE SHAP: FAST SHAP VALUE COMPUTATION FOR TREES

Despite the compelling theoretical advantages of SHAP values, their practical use is hindered by two problems:

- (1) The challenge of estimating $E[f(x) | x_S]$ efficiently.
- (2) The exponential complexity of Equation 2.

Here we focus on tree models and propose fast SHAP value estimation methods specific to trees and ensembles of trees. We start by defining a slow but straightforward algorithm, then present the much faster and more complex Tree SHAP algorithm.

3.1 Estimating SHAP values directly in $O(TL2^M)$ time

If we ignore computational complexity then we can compute the SHAP values for a tree by estimating $E[f(x) | x_S]$ and then using Equation 2 where $f_x(S) = E[f(x) | x_S]$. For a tree model $E[f(x) | x_S]$ can be estimated recursively using Algorithm 1, where v is a vector of node values, which takes the value *internal* for internal nodes. The vectors a and b represent the left and right node indexes for each internal node. The vector t contains the thresholds for each internal node, and d is a vector of indexes of the features used for splitting in internal nodes. The vector r represents the cover of each node (i.e., how many data samples fall in that sub-tree).

The weight w measures what proportion of the training samples matching the conditioning set S fall into each leaf.

Algorithm 1 Estimating $E[f(x) \mid x_S]$

```

procedure EXPVALUE( $x, S, tree = \{v, a, b, t, r, d\}$ )
  procedure G( $j, w$ )
    if  $v_j \neq \text{internal}$  then
      return  $w \cdot v_j$ 
    else
      if  $d_j \in S$  then
        return G( $a_j, w$ ) if  $x_{d_j} \leq t_j$  else G( $b_j, w$ )
      else
        return G( $a_j, wr_{a_j}/r_j$ ) + G( $b_j, wr_{b_j}/r_j$ )
      end if
    end if
  end procedure
return G(1, 1)
end procedure

```

3.2 Estimating SHAP values in $O(TLD^2)$ time

Here we propose a novel algorithm to calculate the same values as above, but in polynomial time instead of exponential time. Specifically, we propose an algorithm that runs in $O(TLD^2)$ time and $O(D^2 + M)$ memory, where for balanced trees the depth becomes $D = \log L$. Recall T is the number of trees, L is the maximum number of leaves in any tree, and M is the number of features.

The intuition of the polynomial time algorithm is to recursively keep track of what proportion of all possible subsets flow down into each of the leaves of the tree. This is similar to running Algorithm 1 simultaneously for all 2^M subsets S in Equation 2. It may seem reasonable to simply keep track of how many subsets (weighted by the cover splitting of Algorithm 1) pass down each branch of the tree. However, this combines subsets of different sizes and so prevents the proper weighting of these subsets, since the weights in Equation 2 depend on $|S|$. To address this we keep track of each possible subset size during the recursion. The *EXTEND* method in Algorithm 2 grows all these subsets according to a given fraction of ones and zeros, while the *UNWIND* method reverses this process and is commutative with *EXTEND*. The *EXTEND* method is used as we descend the tree. The *UNWIND* method is used to undo previous extensions when we split on the same feature twice, and to undo each extension of the path inside a leaf to compute weights for each feature in the path.

In Algorithm 2, m is the path of unique features we have split on so far, and contains four attributes: d the feature index, z the fraction of “zero” paths (where this feature is not in the set S) that flow through this branch, o the fraction of “one” paths (where this feature is in the set S) that flow through this branch, and w which is used to hold the proportion of sets of a given cardinality that are present. We use the dot notation to access these members, and for the whole vector $m.d$ represents a vector of all the feature indexes.

Algorithm 2 reduces the computational complexity of exact SHAP value computation from exponential to low order polynomial for trees and sums of trees (since the SHAP values of a sum of two functions is the sum of the original functions’ SHAP values).

Algorithm 2 Tree SHAP

```

procedure TS( $x, tree = \{v, a, b, t, r, d\}$ )
   $\phi$  = array of  $\text{len}(x)$  zeros
  procedure RECURSE( $j, m, p_z, p_o, p_i$ )
     $m = \text{EXTEND}(m, p_z, p_o, p_i)$ 
    if  $v_j \neq \text{internal}$  then
      for  $i \leftarrow 2$  to  $\text{len}(m)$  do
         $w = \text{sum}(\text{UNWIND}(m, i).w)$ 
         $\phi_{m_i} = \phi_{m_i} + w(m_i.o - m_i.z)v_j$ 
      end for
    else
       $h, c = x_{d_j} \leq t_j ? (a_j, b_j) : (b_j, a_j)$ 
       $i_z = i_o = 1$ 
       $k = \text{FINDFIRST}(m.d, d_j)$ 
      if  $k \neq \text{nothing}$  then
         $i_z, i_o = (m_k.z, m_k.o)$ 
         $m = \text{UNWIND}(m, k)$ 
      end if
      RECURSE( $h, m, i_z r_h / r_j, i_o, d_j$ )
      RECURSE( $c, m, i_z r_c / r_j, 0, d_j$ )
    end if
  end procedure
procedure EXTEND( $m, p_z, p_o, p_i$ )
   $l = \text{len}(m)$ 
   $m = \text{copy}(m)$ 
   $m_{l+1}.(d, z, o, w) = (p_i, p_z, p_o, l = 0 ? 1 : 0)$ 
  for  $i \leftarrow l - 1$  to 1 do
     $m_{i+1}.w = m_{i+1}.w + p_o m_i.w(i/l)$ 
     $m_i.w = p_z m_i.w[(l - i)/l]$ 
  end for
  return  $m$ 
end procedure
procedure UNWIND( $m, i$ )
   $l = \text{len}(m)$ 
   $n = m_l.w$ 
   $m = \text{copy}(m_{1..l-1})$ 
  for  $j \leftarrow l - 1$  to 1 do
    if  $m_i.o \neq 0$  then
       $t = m_j.w$ 
       $m_j.w = n \cdot l / (j \cdot m_i.o)$ 
       $n = t - m_j.w \cdot m_i.z((l - j)/l)$ 
    else
       $m_j.w = (m_j.w \cdot l) / (m_i.z(l - j))$ 
    end if
  end for
  for  $j \leftarrow i$  to  $l - 1$  do
     $m_j.(d, z, o) = m_{j+1}.(d, z, o)$ 
  end for
  return  $m$ 
end procedure
RECURSE(1, [], 1, 1, 0)
return  $\phi$ 
end procedure

```

4 SHAP INTERACTION VALUES

Feature attributions are typically allocated among the input features, one for each feature, but we can gain additional insight by separating *interaction effects* from main effects. If we consider pairwise interactions this leads to a matrix of attribution values representing the impact of all pairs of features on a given model prediction. Since SHAP values are based on classic Shapley values from game theory, a natural extension to interaction effects can be obtained through the more modern Shapley interaction index [12]:

$$\Phi_{i,j} = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!(M - |S| - 2)!}{2(M - 1)!} \nabla_{ij}(S), \quad (3)$$

when $i \neq j$, and

$$\begin{aligned} \nabla_{ij}(S) &= f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S) \\ &= f_x(S \cup \{i,j\}) - f_x(S \cup \{j\}) - [f_x(S \cup \{i\}) - f_x(S)]. \end{aligned} \quad (5)$$

In Equation 3 the SHAP interaction value between feature i and feature j is split equally between each feature so $\Phi_{i,j} = \Phi_{j,i}$ and the total interaction effect is $\Phi_{i,j} + \Phi_{j,i}$. The main effects for a prediction can then be defined as the difference between the SHAP value and the SHAP interaction values for a feature:

$$\Phi_{i,i} = \phi_i - \sum_{j \neq i} \Phi_{i,j}. \quad (6)$$

These SHAP interaction values follow from similar axioms as SHAP values, and allow the separate consideration of main and interaction effects for individual model predictions. This separation can uncover important interactions captured by tree ensembles that might otherwise be missed (Figure 10 in Section 5.5).

While SHAP interaction values can be computed directly from Equation 3, we can leverage Algorithm 2 to drastically reduce their computational cost for tree models. As highlighted in Equation 5 SHAP interaction values can be interpreted as the difference between the SHAP values for feature i when feature j is present and the SHAP values for feature i when feature j is absent. This allows us to use Algorithm 2 twice, once while ignoring feature j as fixed to present, and once with feature j absent. This leads to a run time of $O(TMLD^2)$, since we repeat the process for each feature. Note that even though this computational approach does not seem to directly enforce symmetry, the resulting Φ matrix is always symmetric.

5 EXPERIMENTS AND APPLICATIONS

We compare Tree SHAP and SHAP interaction values with previous methods through both traditional metrics and three new applications we propose for individualized feature attributions: supervised clustering, SHAP summary plots, and SHAP dependence plots.²

5.1 Agreement with Human Intuition

To validate that the SHAP values in Model A of Figure 1 are the most natural assignment of credit we ran a user study to measure people’s intuitive feature attribution values. Model A’s tree was shown to participants and said to represent risk for a certain disease. They were told that when a given person was found to have both a

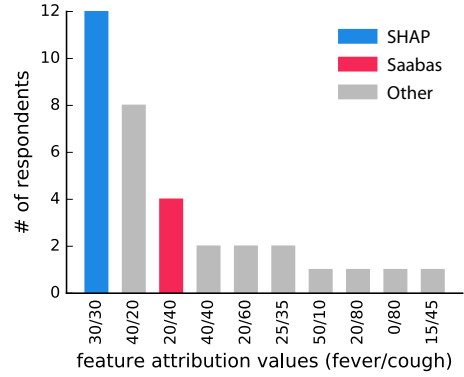


Figure 3: Feature attribution values from 34 participants shown the tree from Model A in Figure 1. The first number represents the allocation to the Fever feature, while the second represents the allocation to the Cough feature. Participants from Amazon Mechanical Turk were not selected for machine learning expertise. No constraints were placed on the feature attribution values users entered.

cough and fever their risk went up from the prior risk of 20 (the expected value of risk) to a risk of 80. Participants were then asked to apportion the 60 point change in risk among the Cough and Fever features as they saw best.

Figure 3 presents the results of the user study for Model A. The equal distribution of credit used by SHAP values was found to be the most intuitive. A smaller number of participants preferred to give greater weight to the first feature to be split on (Fever), while still fewer followed the allocation of the Saabas method and gave greater weight to the second feature split on (Cough).

5.2 Computational Performance

Figure 5 demonstrates the significant run time improvement provided by Algorithm 2. Problems that were previously intractable for exact computation are now inexpensive. An XGBoost model with 1,000 depth 10 trees over 100 input features can now be explained in 0.08 seconds.

5.3 Supervised Clustering

One intriguing application enabled by individualized feature attributions is what we term “supervised clustering,” where instead of using an unsupervised clustering method directly on the data features, you run clustering on the feature attributions.

Supervised clustering naturally handles one of the most challenging problems in unsupervised clustering: determining feature weightings (or equivalently, determining a distance metric). Many times we want to cluster data using features with very different units. Features may be in dollars, meters, unit-less scores, etc. but whenever we use them as dimensions in a single multidimensional space it forces any distance metric to compare the relative importance of a change in different units (such as dollars vs. meters). Even if all our inputs are in the same units, often some features are more

²Jupyter notebooks to compute all results are available at http://github.com/slundberg/shap/notebooks/tree_shap_paper

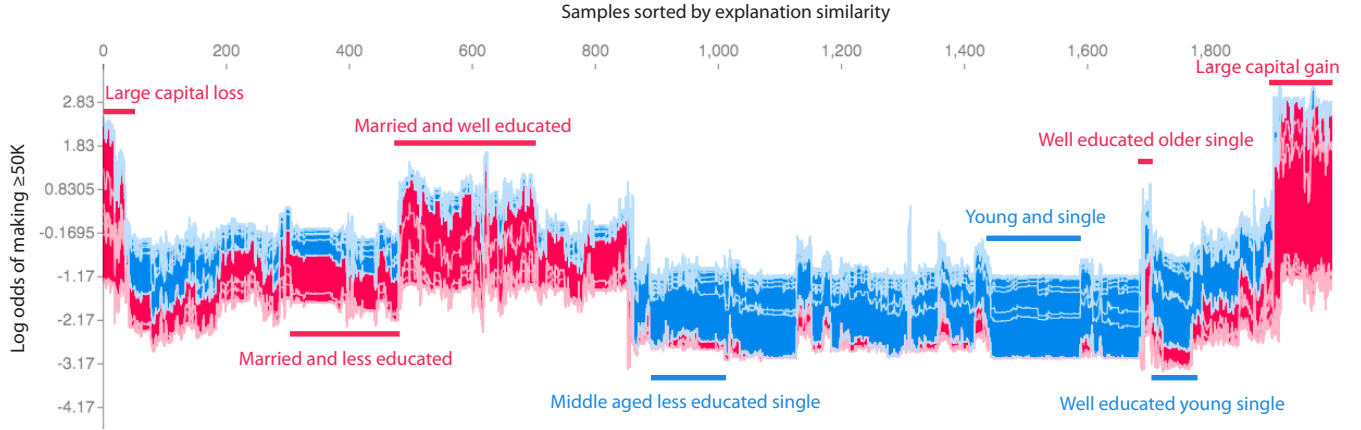


Figure 4: Supervised clustering with SHAP feature attributions in the UCI census dataset identifies among 2,000 individuals distinct subgroups of people that share similar reasons for making money. An XGBoost model with 500 trees of max depth six was trained on demographic data using a shrinkage factor of $\eta = 0.005$. This model was then used to predict the log odds that each person makes $\geq \$50K$. Each prediction was explained using Tree SHAP, and then clustered using hierarchical agglomerative clustering (imagine a dendrogram above the plot joining the samples). Red feature attributions push the score higher, while blue feature attributions push the score lower (as in Figure 2 but rotated 90°). A few of the noticeable subgroups are annotated with the features that define them.

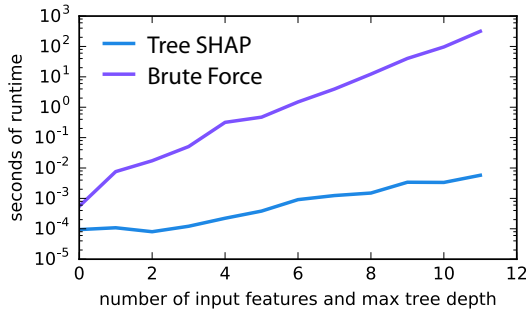


Figure 5: Runtime improvement of Algorithm 2 over using Equation 2 and Algorithm 1. An XGBoost model with 50 trees was trained using an equally increasing number of input features and max tree depths. The time to explain one input vector is reported.

important than others. Supervised clustering uses feature attributions to naturally convert all the input features into values with the same units as the model output. This means that a unit change in any of the feature attributions is comparable to a unit change in any other feature attribution. It also means that fluctuations in the feature values only effect the clustering if those fluctuations have an impact on the outcome of interest.

Here we demonstrate the use of supervised clustering on the classic UCI census dataset [15]. For this dataset the goal is to predict from basic demographic data if a person is likely to make more than \$50K annually. By representing the positive feature attributions as red bars and the negative feature attributions as blue bars (as in Figure 2), we can stack them against each other to visually

represent the model output as their sum. Figure 4 does this vertically for predictions from 2,000 people from the census dataset. The explanations for each person are stacked horizontally according the leaf order of a hierarchical clustering of the SHAP values. This groups people with similar reasons for a predicted outcome together. The formation of distinct subgroups of people demonstrates the power of supervised clustering to identify groups that share common factors related to income level.

One way to quantify the improvement provided by SHAP values over the heuristic Saabas attributions is by examining how well supervised clustering based on each method explains the variance of the model output (note global feature attributions are not considered since they do not enable this type of supervised clustering). If feature attribution values well-represent the model then supervised clustering groups will have similar function outputs. Since hierarchical clusterings encode many possible groupings, we plot in Figure 6 the change in the R^2 value as the number of groups shrinks from one group per sample ($R^2 = 1$) to a single group ($R^2 = 0$). For the census dataset, groupings based on SHAP values outperform those from Saabas values (Figure 6A). For a dataset based on cognitive scores for Alzheimer’s disease SHAP values significantly outperform Saabas values (Figure 6B). This second dataset contains 200 gene expression module levels [4] as features and CERAD cognitive scores as labels [19].

5.4 Identification of Influential Features

Feature attribution values are commonly used to identify which features influenced a model’s prediction the most. To compare methods, the change in a model’s prediction can be computed when the most influential feature is perturbed. Figure 7 shows the result of this experiment on a sentiment analysis model of airline tweets [8]. An XGBoost model with 50 trees of maximum depth 30 was trained

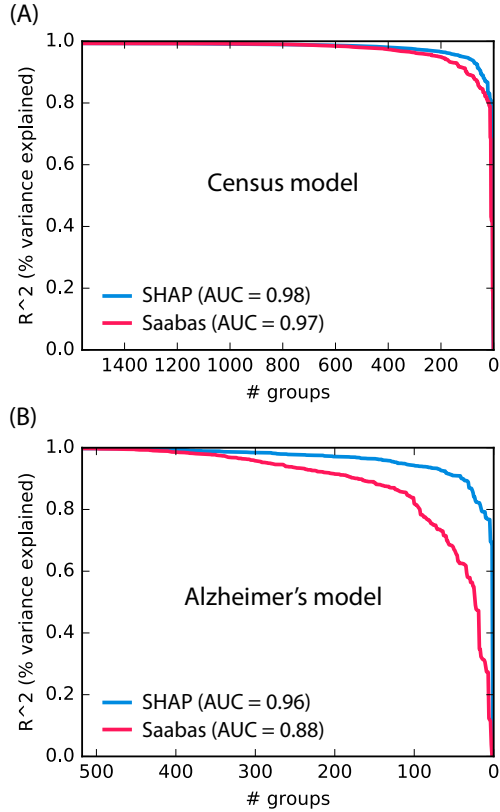


Figure 6: A quantitative measure of supervised clustering performance. If all samples are placed in their own group, and each group predicts the mean value of the group, then the R^2 value (the proportion of model output variance explained) will be 1. If groups are then merged one-by-one the R^2 will decline until when there is only a single group it will be 0. Hierarchical clusterings that well separate the model output value will retain a high R^2 longer during the merging process. Here supervised clustering with SHAP values outperformed the Saabas method in both (A) the census data clustering shown in Figure 4, and (B) a clustering from gene-based predictions of Alzheimer’s cognitive scores.

on 11,712 tweets with 1,686 bag-of-words features. Each tweet had a sentiment score label between -1 (negative) and 1 (positive). The predictions of the XGBoost model were then explained for 2,928 test tweets. For each method we choose the most influential negative feature and replaced it with the value of the same feature in another random tweet from the training set (this is designed to mimic the feature being unknown). The new input is then re-run through the model to produce an updated output. If the chosen feature significantly lowered the model output, then the updated model output should be higher than the original. By tracking the total change in model output as we progress through the test tweets we observe that SHAP values best identify the most influential negative feature. Since global methods only select a single feature for the whole dataset we only replaced this feature when it would likely

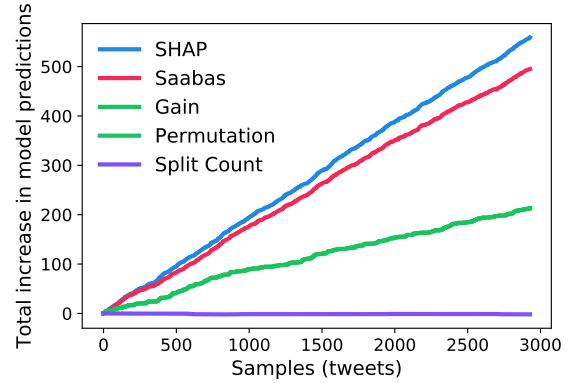


Figure 7: The total increase in a sentiment model’s output when the most negative feature is replaced. Five different attribution methods were used to determine the most negative feature for each sample. The higher the total increase in model output, the more accurate the attribution method was at identifying the most influential negative feature.

increase the sentiment score (for gain and permutation this meant randomly replacing the “thank” feature when it was missing, for split count it was the word “to”).

5.5 SHAP Plots

Plotting the impact of features in a tree ensemble model is typically done with a bar chart to represent global feature importance, or a partial dependence plot to represent the effect of changing a single feature [11]. However, since SHAP values are individualized feature attributions, unique to every prediction, they enable new, richer visual representations. *SHAP summary plots* replace typical bar charts of global feature importance, and *SHAP dependence plots* provide an alternative to partial dependence plots that better capture interaction effects.

To explore these visualizations we trained an XGBoost Cox proportional hazards model on survival data from the classic NHANES I dataset [18] using the NHANES I Epidemiologic Followup Study [7]. After selection for the presence of basic blood test data we obtained data for 9,932 individuals followed for up to 20 years after baseline data collection for mortality. Based on a 80/20 train/test split we chose to use 7,000 trees of maximum depth 3, $\eta = 0.001$, and 50% instance sub-sampling. We then used these parameters and trained on all individuals to generate the final model.

5.5.1 SHAP Summary Plots. Standard feature importance bar charts give a notion of relative importance in the training dataset, but they do not represent the range and distribution of impacts that feature has on the model’s output, and how the feature’s value relates to it’s impact. SHAP summary plots leverage individualized feature attributions to convey all these aspects of a feature’s importance while remaining visually concise (Figure 8). Features are first sorted by their global impact $\sum_{j=1}^N |\phi_i^{(j)}|$, then dots representing the SHAP values $\phi_i^{(j)}$ are plotted horizontally, stacking vertically when they run out of space. This vertical stacking creates an effect similar

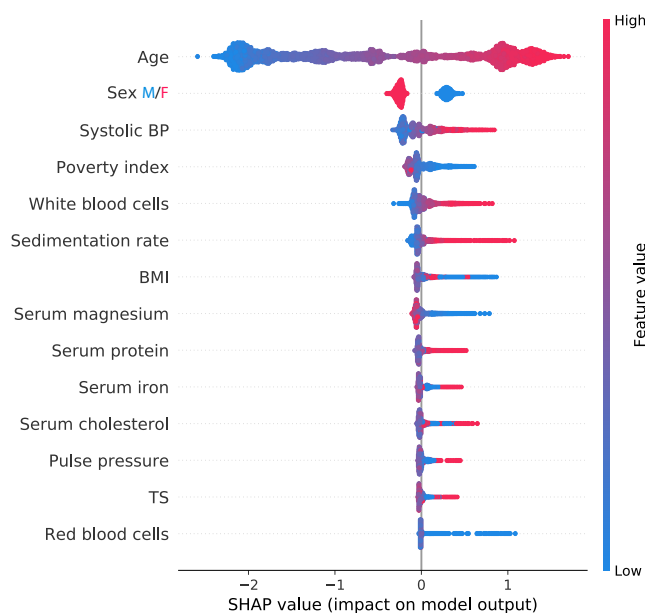


Figure 8: SHAP summary plot of a 14 feature XGBoost survival model on 20 year mortality followup data from NHANES I [18]. The higher the SHAP value of a feature, the higher your log odds of death in this Cox hazards model. Every individual in the dataset is run through the model and a dot is created for each feature attribution value, so one person gets one dot on each feature’s line. Dot’s are colored by the feature’s value for that person and pile up vertically to show density.

to violin plots but without an arbitrary smoothing kernel width. Each dot is colored by the value of that feature, from low (blue) to high (red). If the impact of the feature on the model’s output varies smoothly as its value changes then this coloring will also have a smooth gradation. In Figure 8 we see (unsurprisingly) that age at baseline is the most important risk factor for death over the next 20 years. The density of the age plot shows how common different ages are in the dataset, and the coloring shows a smooth increase in the model’s output (a log odds ratio) as age increases. In contrast to age, systolic blood pressure only has a large impact for a minority of people with high blood pressure. The general trend of long tails reaching to the right, but not to the left, means that extreme values of these measurements can significantly raise your risk of death, but cannot significantly lower your risk.

5.5.2 SHAP Dependence Plots. As described in Equation 10.47 of Friedman et al. (2001), partial dependence plots represent the expected output of a model when the value of a specific variable (or group of variables) is fixed. The values of the fixed variables are varied and the resulting expected model output is plotted. Plotting how the expected output of a function changes as we change a feature helps explain how the model depends on that feature.

SHAP values can be used to create a rich alternative to partial dependence plots, which we term SHAP dependence plots. SHAP

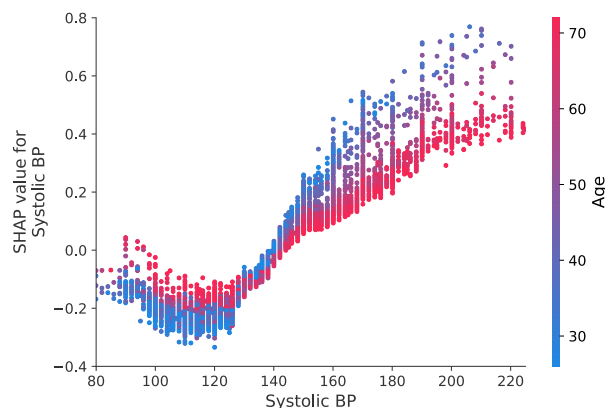


Figure 9: Each dot is a person. The x-axis is their systolic blood pressure and the y-axis is the SHAP value attributed to their systolic blood pressure. Higher SHAP values represent higher risk of death due to systolic blood pressure. Coloring each dot by the person’s age reveals that high blood pressure is more concerning to the model when you are young (this represents an interaction effect).

dependence plots use the SHAP value of a feature for the y-axis and the value of the feature for the x-axis. By plotting these values for many individuals from the dataset we can see how the feature’s attributed importance changes as its value varies (Figure 9). While standard partial dependence plots only produce lines, SHAP dependence plots capture vertical dispersion due to interaction effects in the model. These effects can be visualized by coloring each dot with the value of an interacting feature. In Figure 9 coloring by age shows that high blood pressure is more alarming when you are young. Presumably because it is both less surprising as you age, and possibly because it takes time for high blood pressure to lead to fatal complications.

Combining SHAP dependence plots with SHAP interaction values can reveal global interaction patterns. Figure 10A plots the SHAP main effect value for systolic blood pressure. Since SHAP main effect values represents the impact of systolic blood pressure after all interaction effects have been removed (Equation 6), there is very little vertical dispersion in Figure 10A. Figure 10B shows the SHAP interaction value of systolic blood pressure and age. As suggested by the coloring in Figure 9, this interaction accounts for most of the vertical variance in the systolic blood pressure SHAP values.

6 CONCLUSION

Several common feature attribution methods for tree ensembles are inconsistent, meaning they can lower a feature’s assigned importance when the true impact of that feature actually increases. This can prevent the meaningful comparison of feature attribution values. In contrast, SHAP values consistently attribute feature importance, better align with human intuition, and better recover influential features. By presenting the first polynomial time algorithm for SHAP values in tree ensembles, we make them a practical

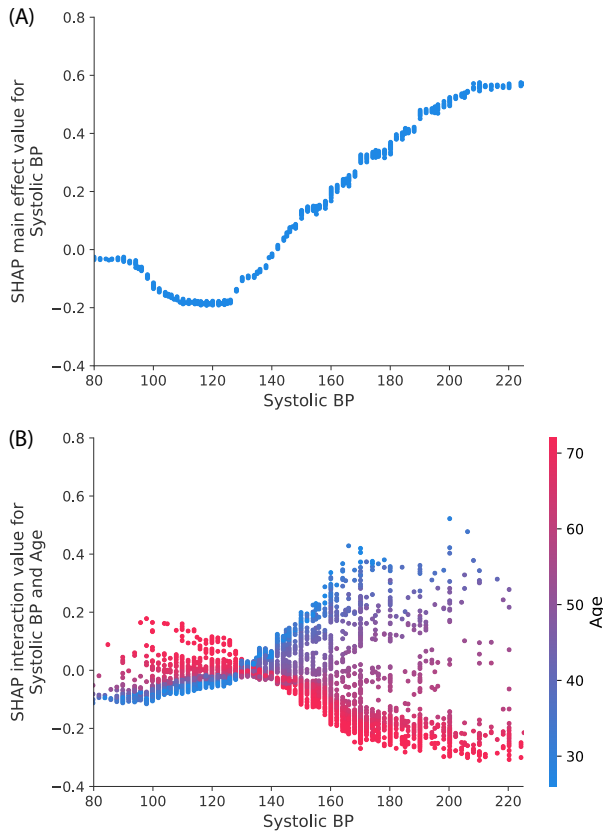


Figure 10: SHAP interaction values separate the impact of systolic blood pressure into main effects (A; Equation 6) and interaction effects (B; Equation 3). Systolic blood pressure has a strong interaction effect with age, so the sum of (A) and (B) nearly equals Figure 9. There is very little vertical dispersion in (A) since all the interaction effects have been removed.

replacement for previous methods. We further defined SHAP interaction values as a consistent way of measuring potentially hidden pairwise interaction relationships. Tree SHAP’s exponential speed improvements open up new practical opportunities, such as supervised clustering, SHAP summary plots, and SHAP dependence plots, that advance our understanding of tree models.

Acknowledgements: Vadim Khotilovich for helpful feedback.

REFERENCES

- [1] Lidia Auret and Chris Aldrich. 2011. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems* 105, 2 (2011), 157–170.
- [2] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Mäzler. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11, Jun (2010), 1803–1831.
- [3] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and regression trees*. CRC press.
- [4] Safiye Celik, Benjamin Logsdon, and Su-In Lee. 2014. Efficient dimensionality reduction for high-dimensional network estimation. In *International Conference on Machine Learning*. 1953–1961.

- [5] S Chebroly, A Abraham, and J Thomas. 2005. Feature deduction and ensemble design of intrusion detection systems. *Computers & security* 24, 4 (2005), 295–307.
- [6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.
- [7] Christine S Cox, Jacob J Feldman, Cordell D Golden, Madelyn A Lane, Jennifer H Madans, Michael E Mussolino, and Sandra T Rothwell. 1997. Plan and operation of the NHANES I Epidemiologic Followup Study, 1992. (1997).
- [8] Crowdfunder. 2015. Twitter US Airline Sentiment. <https://www.kaggle.com/crowdfunder/twitter-airline-sentiment>. (2015). Accessed: 2018-02-06.
- [9] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 598–617.
- [10] R Diaz-Uriarte and S De Andres. 2006. Gene selection and classification of microarray data using random forest. *BMC bioinformatics* 7, 1 (2006), 3.
- [11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin.
- [12] Katsuhige Fujimoto, Ivan Kojadinovic, and Jean-Luc Marichal. 2006. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior* 55, 1 (2006), 72–99.
- [13] A Irrthum, L Wehenkel, P Geurts, et al. 2010. Inferring regulatory networks from expression data using tree-based methods. *PLoS one* 5, 9 (2010), e12776.
- [14] Hemant Ishwaran et al. 2007. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics* 1 (2007), 519–537.
- [15] M. Lichman. 2013. UCI ML Repository. (2013). <http://archive.ics.uci.edu/ml>
- [16] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems* 30. Curran Associates, Inc., 4768–4777. <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- [17] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2017. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. *bioRxiv* (2017), 206540.
- [18] Henry W Miller. 1973. Plan and operation of the health and nutrition examination survey, United States, 1971–1973. *DHEW publication no.(PHS)-Dept. of Health, Education, and Welfare (USA)* (1973).
- [19] Suzanne S Mirra, A Heyman, D McKeel, SM Sumi, Barbara J Crain, LM Brownlee, FS Vogel, JP Hughes, G Van Belle, L Berg, et al. 1991. The Consortium to Establish a Registry for Alzheimer’s Disease (CERAD) Part II. Stand. of the neuropathologic assessments== of Alzheimer’s disease. *Neurology* 41, 4 (1991), 479–479.
- [20] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *JMLR* 12, Oct (2011), 2825–2830.
- [21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*. ACM, 1135–1144.
- [22] Greg Ridgeway. 2010. Generalized boosted regression models. Documentation on the R Package *gbm*, version 1.6–3. (2010).
- [23] W Rodenburg, G Heidema, J Boer, I Bovee-Oudenhoven, E Feskens, E Mariman, and J Keijer. 2008. A framework to identify physiological responses in microarray-based gene expression studies: selection and interpretation of biologically relevant genes. *Physiological genomics* 33, 1 (2008), 78–90.
- [24] Ando Saabas. 2014. Interpreting random forests. <http://blog.datadive.net/interpreting-random-forests/>. (2014). Accessed: 2017-06-15.
- [25] Marco Sandri and Paola Zuccolotto. 2008. A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics* 17, 3 (2008), 611–628.
- [26] C Strobl, A Boulesteix, T Kneib, T Augustin, and A Zeileis. 2008. Conditional variable importance for random forests. *BMC bioinformatics* 9, 1 (2008), 307.
- [27] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41, 3 (2014), 647–665.