# Explaining black box decisions by Shapley cohort refinement

Masayoshi Mase
Hitachi Ltd.

Art B. Owen
Stanford University

Benjamin Seiler
Stanford University

October 2019

**Abstract**

We introduce a variable importance measure to quantify the impact of individual variables to a decision made by a black box function. Our measure is based on the Shapley value from cooperative game theory. Many measures of variable importance operate by changing some predictor values with others held fixed, and that usually creates unlikely or even impossible combinations. Our cohort refinement Shapley approach measures variable importance only using observed data points. Instead of changing the value of a predictor we include or exclude subjects similar to the target subject on that predictor to form a similarity cohort. Then we apply Shapley value to the cohort averages. We also introduce a game theoretic way to aggregate multiple explanations and we illustrate the method on real data sets (titanic and Boston housing).

## 1 Introduction

Black box prediction models used in statistics, machine learning and artificial intelligence have been able to make increasingly accurate predictions, but it remains hard to understand those predictions. See for example, Štrumbelj and Kononenko (2010, 2014), Ribeiro et al. (2016), Sundararajan and Najmi (2019) and the book of Molnar (2018).

Part of understanding predictions is understanding which variables are important. A variable could be important because changing it makes a causal difference, or because changing it makes a large change to our predictions or because leaving it out of a model reduces that model's prediction accuracy (Jiang and Owen, 2003). Importance by one of these criteria need not imply importance by another, though additional assumptions may allow a causal implication to be made from one of the other measures (Pearl, 2009; Zhao and Hastie, 2019). We could be interested in variables that are important overall or in variables that explain one single prediction, such as why a given person was or was not approved for a loan, or why a given patient was or was not placed in an intensive care unit. We use the term impact for the quantitiative change in a

prediction that can be attributed to a variable. This impact can be positive or negative. Importance is then about the absolute value of the impact being large or relatively large.

In this paper, we measure the impact of individual predictor variables used by a model, in order to explain why a given prediction was made. Because we are explaining a given prediction from a given model, we do not address whether that prediction had a sound causal basis. Sound or unsound, we want to understand why it occured, and that understanding might even lead us to conclude that a model is unsound. We also do not consider what the effect of retraining with a different set of predictors would have been, because those differently trained models were not the ones that made the decision.

To fix ideas, we suppose that the decision for a target subject $t$ was based on a vector $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tn})$ of $n$ different predictor variables after training on a data set from $s$ subjects. We will speak of predictors rather than features because features can be constructed as transformations of one or more predictors and our main interest is providing an explanation in terms of the originally measured quantities. While it may be reasonable to make separate attributions for say $x_{tj}$ and $x_{tj}^2$, we leave that for later work. The predictors may be real or categorical among other possibilities. The prediction for subject $i$ is $f(\boldsymbol{x}_i)$ for a function $f$ of a potentially quite complicated form. In the case of a loan, $f(\boldsymbol{x})$ might be a binary variable indicating 1 if the loan should be made to a subject with predictor vector $\boldsymbol{x}$, and 0 otherwise. Or it could be an estimate of the probability that the loan will be repaid, or an estimate of expected return to the lender for making this loan, taking account of administrative costs, default possibilities, the outlook for interest rates, and so on.

Even with the entire function at our disposal in software, it can still be a challenge to quantify a variable's impact. There may be numerous combinations of counterfactual predictors $\boldsymbol{x}$ that could have changed the prediction. The problem of computing the importance of inputs to a function comes up frequently in global sensitivity analysis (Saltelli et al., 2008). Then pick-freeze methods that change some but not all components of $\boldsymbol{x}$ and track how $f$ changes are the norm (Sobol', 1993; Gamboa et al., 2016). There, one usually assumes that the $n$ input variables are statistically independent of each other, and even then the problem is challenging. Black box prediction functions are usually fit to predictors related by complicated dependence patterns, and then predictor independence is extremely unrealistic. Changing some predictors independently of others can lead to predictor combinations far from anything that has been seen in the training data (e.g., a home with many rooms but few square feet) or even impossible combinations (e.g., birth date after graduation date). Those cases are not ones where we can expect the fitted model to be valuable, causing us to doubt that they belong in the explanation.

Our approach does not use any variable combinations that never arose in the sample. Instead, for each predictor, every subject in the data set is either similar to the target subject or not similar. Ways to define similarity are discussed below. Given $n$ predictors, there are $2^n$ different sets of predictors on which subjects can be similar to the target. We form $2^n$ different cohorts of subjects,

each consisting of subjects similar to the target on a subset of predictors, without regard to whether they are also similar on any of the other predictors. At one extreme is a set of all predictors, and a cohort that is similar to the target in every way. At the other extreme, the empty predictor set yields the set of all subjects.

We can refine the grand cohort of all subjects towards the target subject by removing subjects that mismatch the target on one or more predictors. The predictors that change the cohort mean the most when we restrict to similar subjects, are the ones that we take to be the most important in explaining why the target subject's prediction is different from that of the other subjects.

We will define the impact of a variable through the Shapley value. Shapley value has been used in model explanation for machine learning (Štrumbelj and Kononenko, 2010, 2014; Lundberg and Lee, 2017; Sundararajan and Najmi, 2019) and for computer experiments (Owen, 2014; Song et al., 2016; Owen and Prieur, 2017). See Sundararajan and Najmi (2019) for a survey. We will present Shapley value before defining our measures. We call this approach cohort refinement Shapley, or cohort Shapley (CS) for short.

The closest method to our proposal is baseline Shapley from Sundararajan and Najmi (2019). Baseline Shapley compares the predictions for a target subject $t$ with predictors $\boldsymbol{x}_t$ to the predictions from a baseline predictor vector $\boldsymbol{x}_b$. There is not necessarily a subject whose predictors are $\boldsymbol{x}_b$. We can make changes to some predictors $\boldsymbol{x}_{b,j}$ replacing them by the corresponding values $\boldsymbol{x}_{t,j}$ from the target subject $t$, and recording how $f$ changes. Baseline Shapley can construct and use improbable or even impossible combinations of predictors, as the authors note, while cohort Shapley does not.

Sundararajan and Najmi (2019) mention a second problem with baseline Shapley. It arises when two predictors are highly correlated. Consider an extreme case where $x_{i,j} = x_{i,k}$ for all subjects $i$ and two different predictors $j$ and $k$. The prediction function $f$ might use these two predictors equally or it might make an arbitrary choice to use one and completely ignore the other, or the precise combination could be in between these extremes in some very complicated way. The importance of predictors $j$ and $k$ from baseline Shapley will then depend on those choices because they affect the value that $f$ will take on a hypothetical point where $x_{i,j} \neq x_{i,k}$ holds. For CS, let us assume that for any subject $i$, and two equivalent predictors $j$ and $k$, we will have $x_{i,j}$ similar to $x_{t,j}$ if and only if $x_{i,k}$ is similar to $x_{t,k}$. In that case we will find that predictors $j$ and $k$ get equal Shapley values, even if the model ignores one of them.

This paper is organized as follows. Section 2 gives our notation, and reviews Shapley value, the functional ANOVA decomposition, and the anchored decomposition. Section 3 defines similarity and similarity-based cohorts, with a small example to illustrate those sets. Section 4 presents cohort Shapley importance measures. Section 5 describes a game theoretic way to aggregate impacts over a set of target subjects, such as all subjects in the data set. Section 6 shows cohort Shapley on some real data sets. Section 7 discusses strengths and weaknesses of cohort Shapley and also how it addresses a different goal than baseline Shapley does.

# 2 Notation and background

The predictor vector for subject $i$ is $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})$ where $n$ is the number of predictors in the model. Each $x_{ij}$ belongs to a set $X_j$ which may consist of real or binary variables or some other types. There is a black box function $f(\boldsymbol{x}) \in \mathbb{R}$ that is used to predict an outcome for a subject with predictor vector $\boldsymbol{x}$. We write $y = f(\boldsymbol{x})$ and $y_i = f(\boldsymbol{x}_i)$. There is a target subject $t$ and we would like an explanation about which predictors $x_{tj}$ are the most important determinants of $y_t = f(\boldsymbol{x}_t)$. We assume that $t$ is in the set of subjects with available data, although this subject might not have been used in training the model.

The set $\{1, 2, \ldots, n\}$ is denoted by $1{:}n$. We will need to manipulate subsets of $1{:}n$. For $u \subseteq 1{:}n$ we let $|u|$ be its cardinality. The complementary set $1{:}n \setminus u$ is denoted by $-u$, especially in subscripts. Sometimes a point must be created by combining parts of two other points. The point $\boldsymbol{y} = \boldsymbol{x}_u{:}\boldsymbol{z}_{-u}$ has $y_j = x_j$ for $j \in u$ and $y_j = z_j$ for $j \notin u$. Furthermore, we sometimes use $j$ and $-j$ in place of the more cumbersome $\{j\}$ and $-\{j\}$. For instance, $\boldsymbol{x}_{-j}{:}\boldsymbol{z}_j$ is what we get by replacing the $j$'th input to $\boldsymbol{x}$ by $z_j$.

## 2.1 Shapley value

Shapley value (Shapley, 1952) is used in game theory to define a fair allocation of rewards to a team that has cooperated to produce something of value. Suppose that a team of $n$ people produce a value $\mathrm{val}(1{:}n)$, and that we have at our disposal the value $\mathrm{val}(u)$ that would have been produced by the team $u$, for all $2^n$ teams $u \subseteq 1{:}n$, including $\mathrm{val}(\varnothing) = 0$. Let $\phi_j$ be the reward for player $j$.

Shapley introduced quite reasonable criteria:

1) Efficiency: $\sum_{j=1}^{n} \phi_j = \mathrm{val}(1{:}n)$.
2) Symmetry: If $\mathrm{val}(u + i) = \mathrm{val}(u + j)$ for all $u \subseteq 1{:}n \setminus \{i, j\}$, then $\phi_i = \phi_j$.
3) Dummy: if $\mathrm{val}(u + j) = \mathrm{val}(u)$ for all $u \subseteq 1{:}n \setminus \{j\}$, then $\phi_j = 0$.
4) Additivity: if $\mathrm{val}(u)$ and $\mathrm{val}'(u)$ lead to values $\phi_j$ and $\phi'_j$ then the game producing $\mathrm{val} + \mathrm{val}'$ has values $\phi + \phi'$.

He found that the unique valuation that satisfies all four of these criteria is

$$\phi_j = \frac{1}{n} \sum_{u \subseteq -j} \binom{n-1}{|u|}^{-1} (\mathrm{val}(u + j) - \mathrm{val}(u)). \tag{1}$$

Formula (1) is not very intuitive. Another way to explain Shapley value is as follows. We could build a team from $\varnothing$ to $1{:}n$ in $n$ steps, adding one member at a time. There are $n!$ different orders in which to add team members. The Shapley value $\phi_j$ is the increase in value coming from the addition of member $j$, averaged over those $n!$ different orders.

## 2.2 Function decompositions

Function decompositions, also called high dimensional model representations (HDMR), write a function of $n$ inputs as a sum of functions, each of which

depend only on one of the $2^n$ subsets of inputs. Because $f$ and $\boldsymbol{x}$ have other uses in this paper we present the decomposition for $g(\boldsymbol{z})$. Let $g$ be a function of $\boldsymbol{z} = (z_1, \ldots, z_n)$ with $z_j \in Z_j$. In these decompositions we write

$$g(\boldsymbol{z}) = \sum_{u \subseteq 1:n} g_u(\boldsymbol{z})$$

where $g_u(\boldsymbol{z})$ depends on $\boldsymbol{z}$ only through $\boldsymbol{z}_u$. Many such decompositions are possible (Kuo et al., 2010).

The best known decomposition is the analysis of variance (ANOVA) decomposition. It applies to random $\boldsymbol{z}$ with independent components $z_j \in Z_j$. If $\mathbb{E}(g(\boldsymbol{z})^2) < \infty$, then we write

$$\begin{aligned} g_\varnothing(\boldsymbol{z}) &= \mu \equiv \mathbb{E}(g(\boldsymbol{z})), \quad \text{followed by} \\ g_u(\boldsymbol{z}) &= \mathbb{E}\biggl(g(\boldsymbol{z}) - \sum_{v \subsetneq u} g_v(\boldsymbol{z}) \;\Big|\; \boldsymbol{z}_u\biggr) = \mathbb{E}(g(\boldsymbol{z}) \,|\, \boldsymbol{z}_u) - \sum_{v \subsetneq u} g_v(\boldsymbol{z}), \end{aligned}$$

for non-empty $u \subseteq 1{:}n$. The effects $g_u$ are mutually orthogonal in that for subsets $u \neq v$, we have $\mathbb{E}(g_u(\boldsymbol{z})g_v(\boldsymbol{z})) = 0$. Letting $\sigma_u^2 = \mathrm{var}(g_u(\boldsymbol{z}))$, it follows from orthogonality that

$$\sigma^2(g) \equiv \mathrm{var}(g(\boldsymbol{z})) = \sum_{u \subseteq 1:n} \sigma_u^2(g).$$

We can recover effects from conditional expectations, via inclusion-exclusion,

$$g_u(\boldsymbol{z}) = \sum_{v \subseteq u} (-1)^{|u-v|} \mathbb{E}(g(\boldsymbol{z}) \,|\, \boldsymbol{z}_v). \tag{2}$$

See Owen (2013) for history and derivations of this functional ANOVA.

We will need the anchored decomposition, which goes back at least to Sobol' (1969). It is also called cut-HDMR (Aliş and Rabitz, 2001) in chemistry, and finite differences-HDMR in global sensitivity analysis (Sobol', 2003). We begin by picking a reference point $\boldsymbol{c}$ called the anchor, with $c_j \in Z_j$ for $j = 1, \ldots, n$. The anchored decomposition is

$$\begin{aligned} g(\boldsymbol{z}) &= \sum_{u \subseteq 1:n} g_{u,\boldsymbol{c}}(\boldsymbol{z}), \quad \text{with} \\ g_{\varnothing,\boldsymbol{c}}(\boldsymbol{z}) &= g(\boldsymbol{c}), \quad \text{and} \\ g_{u,\boldsymbol{c}}(\boldsymbol{z}) &= g(\boldsymbol{z}_u{:}\boldsymbol{c}_{-u}) - \sum_{v \subsetneq u} g_{v,\boldsymbol{c}}(\boldsymbol{z}). \end{aligned}$$

We have replaced averaging over $\boldsymbol{z}_{-u}$ by plugging in the anchor value via $\boldsymbol{z}_{-u} = \boldsymbol{c}_{-u}$. If $j \in u$ and $z_j = c_j$, then $g(\boldsymbol{z})_{u,\boldsymbol{c}} = 0$. We do not need independence of the $z_j$, or even randomness for them and we do not need mean squares. What we need is that when $g(\boldsymbol{z})$ is defined, so is $g(\boldsymbol{z}_u{:}\boldsymbol{c}_{-u})$ for any $u \subseteq 1{:}n$.

The main effect in an anchored decomposition is $g_{j,\boldsymbol{c}}(\boldsymbol{z}) = g(\boldsymbol{z}_j{:}\boldsymbol{c}_{-j}) - g(\boldsymbol{c})$ and the two factor term for indices $j \neq k$ is

$$\begin{aligned} g_{\{j,k\},\boldsymbol{c}}(\boldsymbol{z}) &= g(\boldsymbol{z}_{\{j,k\}}{:}\boldsymbol{c}_{-\{j,k\}}) - g_{j,\boldsymbol{c}}(\boldsymbol{z}) - g_{k,\boldsymbol{c}}(\boldsymbol{z}) - g_{\varnothing}(\boldsymbol{z}) \\ &= g(\boldsymbol{z}_{\{j,k\}}{:}\boldsymbol{c}_{-\{j,k\}}) - g(\boldsymbol{z}_j{:}c_{-j}) - g(\boldsymbol{z}_k{:}c_{-k}) + g(\boldsymbol{c}). \end{aligned}$$

For instance if $n = 3$ and $\boldsymbol{c} = \boldsymbol{0}$, then

$$g(z_1, z_2, z_3) = g(z_1, z_2, 0) - g(z_1, 0, 0) - g(0, z_2, 0) + g(0, 0, 0).$$

The version of (2) for the anchored decomposition is

$$g_{u,\boldsymbol{c}}(\boldsymbol{z}) = \sum_{v \subseteq u} (-1)^{|u-v|} g(\boldsymbol{z}_v{:}\boldsymbol{c}_{-v}),$$

as shown by Kuo et al. (2010).

## 2.3 Shapley for function decompositions

To get a Shapley value for predictor variables, we must first define the value produced by a subset of them. The approach of Štrumbelj and Kononenko (2010) begins with a vector $\boldsymbol{x}$ of independent random predictors from some distribution $F$. They used independent predictors uniformly distributed over finite discrete sets but they could as well be countable or continuous and non-uniform, so long as they are independent. For a target subject $t$, let $f(\boldsymbol{x}_t)$ be the prediction for that subject. They define the value of the predictor set $u \subseteq 1{:}n$ by

$$\Delta(u) = \mathbb{E}(f(\boldsymbol{x}_{t,u}{:}\boldsymbol{z}_{-u})) - \mathbb{E}(f(\boldsymbol{z}))$$

with expectations taken under $\boldsymbol{z} \sim F$. In words, $\Delta(u)$ is the expected change in our predictions at a random point $\boldsymbol{z}$ that comes from specifying that $z_j = x_{t,j}$ for $j \in u$, while leaving $z_j$ random for $j \notin u$.

In their formulation, the total value to be explained is

$$\Delta(1{:}n) = f(\boldsymbol{x}_t) - \mathbb{E}(f(\boldsymbol{z})),$$

the extent to which $f(\boldsymbol{x}_t)$ differs from a hypothetical average prediction over independent predictors. The subset $u$ explains $\Delta(u)$, and from that they derive Shapley value. They define quantities $I(u)$ via $I(\varnothing) = 0$ and

$$\Delta(u) = \sum_{v \subseteq u} I(u).$$

They prove that the Shapley value for predictor $j$ is

$$\phi_j = \sum_{u \subseteq 1:n,\, j \in u} \frac{I(u)}{|u|}.$$

We give a proof of this in Section 4, different from theirs, making use of the anchored decomposition. While their $\Delta(u)$ is defined via expectations of independent random variables, their Shapley value comes via the anchored decomposition applied to those expectations.

A second approach to Shapley value for the ANOVA is to define the value of the set $u$ to be the variance explained by those predictors, $\mathrm{var}(\mathbb{E}(g(\boldsymbol{x})\,|\,\boldsymbol{x}_u))$. With this definition, the Shapley value for $j$ is

$$\phi_j = \sum_{u \subseteq 1:n,\, j \in u} \frac{\sigma_u^2}{|u|}. \tag{3}$$

See Owen (2014). For Shapley value based on variance explained by dependent inputs, see Song et al. (2016) and Owen and Prieur (2017).

# 3 Similarity-based cohorts

A cohort is a set of subjects. For the target subject $t$, we will define a suite of cohorts consisting of subjects similar to $t$ in various ways. The subject $t$ will be in all of those cohorts. First we describe similarity.

## 3.1 Similarity

For each predictor $j$, we define a target-specific similarity function $z_{tj} : X_j \to \{0, 1\}$. If $z_{tj}(x_{ij}) = 1$, then subject $i$ is considered to be similar to subject $t$ as measured by predictor $j$. Otherwise $z_{tj}(x_{ij}) = 0$ means that subject $i$ is dissimilar to subject $t$ for predictor $j$. The simplest similarity is identity:

$$z_{tj}(x_{ij}) = \begin{cases} 1, & x_{ij} = x_{tj}, \\ 0, & \text{else}, \end{cases}$$

which is reasonable for binary predictors or those taking a small number of levels. For real-valued predictors, there may be no $i \neq t$ with $x_{ij} = x_{tj}$ and then we might instead take

$$z_{tj}(x_{ij}) = \begin{cases} 1, & |x_{ij} - x_{tj}| \leqslant \delta_{tj}, \\ 0, & \text{else}, \end{cases}$$

where subject matter experts have chosen $\delta_{tj}$. Taking $\delta_{tj} = 0$ recovers the identity measure of similarity. The two similarity measures above generate an equivalence relation on $X_j$, if $\delta_{tj}$ does not depend on $t$. In general, we do not need $z_{tj}$ to be an equivalence. For instance, we do not need $z_{tj}(x_{ij}) = z_{ij}(x_{tj})$ and would not necessarily have that if we used relative distance to define similarity, via

$$z_{tj}(x_{ij}) = \begin{cases} 1, & |x_{ij} - x_{tj}| \leqslant \delta_j |x_{tj}|, \\ 0, & \text{else}. \end{cases}$$

Table 1: A toy data set of 8 subjects. For each of 3 predictors, $z_{8,j}$ indicates whether a subject is similar to target subject $t = 8$ on predictor $j$.

| Subj | $z_{8,1}$ | $z_{8,2}$ | $z_{8,3}$ |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 |
| 3 | 0 | 1 | 0 |
| 4 | 0 | 1 | 1 |
| 5 | 1 | 0 | 0 |
| 6 | 1 | 0 | 1 |
| 7 | 1 | 1 | 0 |
| 8 | 1 | 1 | 1 |

### 3.2 Cohorts of $t$

We use $1{:}s = \{1, 2, \ldots, s\}$ to define our set of subjects. Let

$$C_{t,u} = \{i \in 1{:}s \mid z_{tj}(x_{ij}) = 1, \quad \text{for all } j \in u\},$$

with $C_{t,\varnothing} = 1{:}s$ by convention. Then $C_{t,u}$ is the cohort of subjects that are similar to the target subject for all predictors $j \in u$ but not necessarily similar for any predictors $j \notin u$. These cohorts are never empty, because we always have $t \in C_{t,u}$. We write $|C_{t,u}|$ for the cardinality of the cohort.

Tables 1 and 2 show the cohort structure for a toy dataset with 8 subjects, three predictors and a target subject $t = 8$. As the cardinality of $u$ increases, the cohort $C_{t,u}$ focusses in on the target subject. The 8 subjects listed there could be generalized to 8 groups of subjects who either match or don't match a target subject from group 8 for those three predictors. Then the cohorts would be unions of those groups. For instance, the cohort $\{6, 8\}$ in Table 1 would become the union of groups 6 and 8. Even if one or more of those groups were empty, none of the cohorts would be empty, again due to subject $t$.

Given a set of cohorts, we define cohort averages

$$\bar{y}_{t,u} = \frac{1}{|C_{t,u}|} \sum_{i \in C_{t,u}} y_i.$$

Then the value of set $u$ is

$$\mathrm{val}_{\mathrm{CS}}(u) = \bar{y}_{t,u} - \bar{y}_{t,\varnothing} = \bar{y}_{t,u} - \bar{y},$$

where $\bar{y} = (1/s)\sum_{i=1}^{s} y_i$. The last equality follows because the cohort with $u = \varnothing$ is the whole data set. The total value to be explained is

$$\mathrm{val}_{\mathrm{CS}}(1{:}n) = \bar{y}_{t,1:n} - \bar{y}.$$

It may well happen that $C_{t,1:n}$ is the singleton $\{t\}$. In that case the total value to be explained is $f(\boldsymbol{x}_t) - \bar{y}$. In this and other settings some of the $\bar{y}_{t,u}$ may be the average of a very small number of subjects' predictions, and potentially poorly determined. We return to this point in Section 7.

Table 2: The $2^3 = 8$ cohorts corresponding to sets $u$ of predictors shown in Table 1. To belong to the cohort for set $u$, a subject must be similar to the target subject for all predictors $j \in u$.

| Set $u$ | Cohort $C_{8,u}$ |
|---|---|
| $\varnothing$ | $\{1,2,3,4,5,6,7,8\}$ |
| $\{1\}$ | $\{5,6,7,8\}$ |
| $\{2\}$ | $\{3,4,7,8\}$ |
| $\{3\}$ | $\{2,4,6,8\}$ |
| $\{1,2\}$ | $\{7,8\}$ |
| $\{1,3\}$ | $\{6,8\}$ |
| $\{2,3\}$ | $\{4,8\}$ |
| $\{1,2,3\}$ | $\{8\}$ |

# 4 Importance measures

For Shapley value, every variable is either 'in or out', and so binary variables underly the approach. Here we compute Shapley values based on function decompositions of a function $g$ defined on $\{0,1\}^n$. The $2^n$ values of that function might themselves be expectations, like the cohort mean in cohort Shapley or the quantity $\Delta$ in the approach of Štrumbelj and Kononenko (2010), but for our purposes here they are just $2^n$ numbers.

When the target point $\boldsymbol{x}_t$ changes, then the Shapley value changes too. Sundararajan and Najmi (2019) consider the effects of continuously varying the target point and describe some invariance and monotonicity properties. For any fixed target $\boldsymbol{x}_t$ and baseline $\boldsymbol{x}_b$, baseline Shapley is defined in terms of the binary variables we consider here.

We use $\boldsymbol{e}_j$ to represent the binary vector of length $n$ with a one in position $j$ and zeroes elsewhere. This is the $j$'th standard basis vector. We then generalize it to $\boldsymbol{e}_u = \mathbf{1}_u{:}\mathbf{0}_{-u}$ for $u \subseteq 1{:}n$. An arbitrary point in $\{0,1\}^n$ is denoted by $\boldsymbol{z}$.

Let $g$ be a function on $\{0,1\}^n$. In our applications, the total value to be explained is $g(\mathbf{1}) - g(\mathbf{0})$, with $\mathbf{1}$ corresponding to matching the target in all $n$ ways and $\mathbf{0}$ corresponding to no matches at all. The value contributed by $u \subseteq 1{:}n$ is $g(\boldsymbol{e}_u) - g(\mathbf{0})$.

## 4.1 Shapley value via anchored decomposition on $\{0,1\}^n$

Because we use the anchored decomposition for functions on $\{0,1\}^n$ instead of the ANOVA, we do not need to define a distribution for $\boldsymbol{z}$. The anchored decomposition on $\{0,1\}^n$ with anchor $\boldsymbol{c} = \mathbf{0}$ has a simple structure.

**Lemma 1.** *For integer $n \geqslant 1$, let $g : \{0,1\}^n$ have the anchored decomposition $g(\boldsymbol{z}) = \sum_{u \subseteq 1:n} g_{u,\mathbf{0}}(\boldsymbol{z})$ with anchor $\mathbf{0}$. Then*

$$g_{u,\mathbf{0}}(\boldsymbol{e}_w) = g_{u,\mathbf{0}}(\mathbf{1}) 1_{u \subseteq w}, \tag{4}$$

*where* $\boldsymbol{e}_w = \mathbf{1}_w{:}\mathbf{0}_{-w}$.

*Proof.* The inclusion-exclusion formula for the binary anchored decomposition is

$$g_{u,\mathbf{0}}(\boldsymbol{z}) = \sum_{v\subseteq u}(-1)^{|u-v|}g(\boldsymbol{z}_v{:}\mathbf{0}_{-v}).$$

Suppose that $z_j = 0$ for $j \in u$. Then, splitting up the alternating sum

$$g_{u,\mathbf{0}}(\boldsymbol{z}) = \sum_{v\subseteq u-j}(-1)^{|u-v|}(g(\boldsymbol{z}_v{:}\mathbf{0}_{-v}) - g(\boldsymbol{z}_{v+j}{:}\mathbf{0}_{-v-j})) = 0$$

because $\boldsymbol{z}_v{:}\mathbf{0}_{-v}$ and $\boldsymbol{z}_{v+j}{:}\mathbf{0}_{-v-j}$ are the same point when $z_j = 0$. It follows that $g_{u,\mathbf{0}}(\boldsymbol{e}_w) = 0$ if $u \subseteq w$ does not hold.

Now suppose that $u \subseteq w$. First $g_{u,\mathbf{0}}(\boldsymbol{z}) = g_{u,\mathbf{0}}(\boldsymbol{z}_u{:}\mathbf{1}_{-u})$ because $g_{u,\mathbf{0}}$ only depends on $\boldsymbol{z}$ through $\boldsymbol{z}_u$. From $u \subseteq w$ we have $(\boldsymbol{e}_w)_u = \mathbf{1}_u$. Then $g_{u,\mathbf{0}}(\boldsymbol{e}_w) = g_{u,\mathbf{0}}(\mathbf{1}_u{:}\mathbf{1}_{-u}) = g_{u,\mathbf{0}}(\mathbf{1})$, completing the proof. □

Now we find the Shapley value for a function on $\{0,1\}^n$ in an anchored decomposition. Štrumbelj and Kononenko (2010) proved this earlier using different methods.

**Theorem 1.** *Let $g(\boldsymbol{z})$ have the anchored decomposition with terms $g_{u,\mathbf{0}}(\boldsymbol{z})$ for $\boldsymbol{z} \in \{0,1\}^n$. Let the set $u \subseteq 1{:}n$ contribute value $g(\boldsymbol{e}_u) - g(\mathbf{0})$. Then the total value is $g(\mathbf{1}) - g(\mathbf{0})$, and the Shapley value for variable $j \in 1{:}n$ is*

$$\phi_j = \sum_{u,\,j\in u}\frac{g_{u,\mathbf{0}}(\mathbf{1}) - g_{u,\mathbf{0}}(\mathbf{0})}{|u|} = \sum_{u,\,j\in u}\frac{g_{u,\mathbf{0}}(\mathbf{1})}{|u|}. \tag{5}$$

*Proof.* For $u \neq \varnothing$, $g_{u,\mathbf{0}}(\mathbf{0}) = 0$, and so the two expressions for $\phi_j$ in (5) are equal. From the definition of Shapley value,

$$\begin{aligned}
\phi_j &= \frac{1}{n}\sum_{v\subseteq -j}\binom{n-1}{|v|}^{-1}(g(\boldsymbol{e}_{v+j}) - g(\mathbf{0})) - (g(\boldsymbol{e}_v) - g(\mathbf{0}))\\
&= \frac{1}{n}\sum_{v\subseteq -j}\binom{n-1}{|v|}^{-1}(g(\boldsymbol{e}_{v+j}) - g(\boldsymbol{e}_v)) \qquad (6)\\
&= \frac{1}{n}\sum_{v\subseteq -j}\binom{n-1}{|v|}^{-1}\sum_{u\subseteq 1:n}(g_{u,\mathbf{0}}(\boldsymbol{e}_{v+j}) - g_{u,\mathbf{0}}(\boldsymbol{e}_v)).
\end{aligned}$$

By Lemma 1,

$$\begin{aligned}
\phi_j &= \frac{1}{n}\sum_{v\subseteq -j}\binom{n-1}{|v|}^{-1}\sum_{u\subseteq 1:n}(g_{u,\mathbf{0}}(\mathbf{1})1_{u\subseteq v+j} - g_{u,\mathbf{0}}(\mathbf{1})1_{u\subseteq v})\\
&= \frac{1}{n}\sum_{u\subseteq 1:n}g_u(\mathbf{1})\sum_{v\subseteq -j}\binom{n-1}{|v|}^{-1}(1_{u\subseteq v+j} - 1_{u\subseteq v}).
\end{aligned}$$

Now

$$1_{u\subseteq v+j} - 1_{u\subseteq v} = 1_{j\in u}1_{j\notin v}1_{v\supseteq u-j}. \tag{7}$$

The cardinality of $v$ for which (7) is nonzero ranges from $|u|-1$ to $n-1$ and so

$$\begin{aligned}\phi_j &= \frac{1}{n}\sum_{u,\,j\in u} g_{u,\mathbf{0}}(\mathbf{1}) \sum_{r=|u|-1}^{n-1} \binom{n-1}{r}^{-1} \sum_{v\subseteq -j} 1_{j\in u}1_{j\notin v}1_{v\supseteq u-j}1_{|v|=r} \\ &= \frac{1}{n}\sum_{u,\,j\in u} g_{u,\mathbf{0}}(\mathbf{1}) \sum_{r=|u|-1}^{n-1} \binom{n-1}{r}^{-1} \binom{n-|u|}{r-|u|+1},\end{aligned}$$

because $v$ contains $u-j$ and $r-|u|+1$ additional indices from $-u$. Simplifying

$$\binom{n-1}{r}^{-1}\binom{n-|u|}{r-|u|+1} = \binom{r}{|u|-1}\binom{n-1}{|u|-1}^{-1}$$

and

$$\sum_{r=|u|-1}^{n-1}\binom{r}{|u|-1} = \binom{n}{|u|}$$

by the "hockey-stick identity". Therefore

$$\phi_j = \frac{1}{n}\sum_{u,\,j\in u} g_{u,\mathbf{0}}(\mathbf{1})\binom{n}{|u|}\binom{n-1}{|u|-1}^{-1} = \sum_{u,\,j\in u}\frac{g_{u,\mathbf{0}}(\mathbf{1})}{|u|}. \qquad \square$$

The right hand side of (5) appears like it might not sum to $g(\mathbf{1})-g(\mathbf{0})$. To verify that it does, write

$$\sum_{j=1}^{n}\sum_{j,\,j\in u}\frac{g_{u,\mathbf{0}}(\mathbf{1})}{|u|} = \sum_{u\neq\varnothing} g_{u,\mathbf{0}}(\mathbf{1}) = -g_{\varnothing,\mathbf{0}}(\mathbf{1}) + \sum_{u\subseteq 1:n} g_{u,\mathbf{0}}(\mathbf{1}) = g(\mathbf{1}) - g(\mathbf{0}).$$

The proof in Štrumbelj and Kononenko (2010) proceeds by substituting the inclusion-exclusion identity into the first expression for $\phi_j$ in (5) and then showing that it is equal to the definition of Shapley value. They also need to explain some of their steps in prose and the version above provides a more 'mechanical' alternative approach.

For $g(\boldsymbol{e}_u) = \mathrm{val}_{\mathrm{CS}}(u) = \bar{y}_{t,u} - \bar{y}$ we get, using inclusion-exclusion

$$\phi_j = \sum_{u,\,j\in u}\frac{1}{|u|}\sum_{v\subseteq u}(-1)^{|u-v|}\bar{y}_{t,v} = \sum_{v\subseteq 1:n}\bar{y}_{t,v}\sum_{u\supseteq v+j}\frac{(-1)^{|u-v|}}{|u|}.$$

This is the Shapley value for any $2^n$ numbers on the corners of $\{0,1\}^n$ provided that $\mathbf{0}$ gets the value 0.

The expression in (6) is easier to interpret. It yields

$$\phi_j = \frac{1}{n}\sum_{v\subseteq -j}\binom{n-1}{|v|}^{-1}(\bar{y}_{t,v+j}-\bar{y}_{t,v}).$$

Here $\bar{y}_{t,v+j}-\bar{y}_{t,v}$ is the difference that refining on variable $j$ makes when we have already refined on the variable set $v$. Now $\phi_j$ is the average over cardinalities $|v|=0,\dots,n-1$ of the average of all $\binom{n-1}{|v|}$ such differences. The contribution from $v=\varnothing$ is given the same weight as the average of all $n-1$ contributions $\{k\}$ to $\{j,k\}$ for $k\neq j$.

# 5 Aggregation

Given a set of per-subject Shapley values, we can explore them graphically and numerically to extract insights. One important task is to compare importance of predictors in aggregate over a set $w\subseteq 1{:}s$ of subjects. While that can be done in numerous ways with summary statistics, such as average absolute Shapley value, we would prefer to derive an aggregate measure from a game so that the aggregate measure that satisfies the four Shapley criteria from Section 2.1.

Because Shapley value is additive over games, we could simply sum the per-subject Shapley values, that we now denote by $\phi_{j,t}$. That will provide an unfortunate cancellation that we seek to avoid. To see the cancellation, suppose that predictor $x_j$ takes the values 0 or 1 and the value $x_j=1$ generally leads to a larger outcome $y=f(\boldsymbol{x})$. We will then tend to get positive cohort Shapley values $\phi_{j,t}$ when $x_{t,j}=1$ and negative ones otherwise. These effects will tend to cancel in $\sum_{t\in w}\phi_{j,t}$, obscuring the impact of $x_j$.

To avoid this cancellation we let

$$\mathrm{sgn}_t = \mathrm{sgn}(f(\boldsymbol{x}_t)-\bar{y}) = \begin{cases} 1, & f(\boldsymbol{x}_t)>\bar{y}\\ 0, & f(\boldsymbol{x}_t)=\bar{y}\\ -1, & f(\boldsymbol{x}_t)<\bar{y},\end{cases}$$

and define the value of set $u$ to be

$$\mathrm{val}^w_{\mathrm{CS}}(u)=\sum_{t\in w}\mathrm{sgn}_t\times(\bar{y}_{t,u}-\bar{y}).$$

This corresponds to a game with total value

$$\mathrm{val}^w_{\mathrm{CS}}(1{:}n)=\sum_{t\in w}\mathrm{sgn}_t\times(\bar{y}_{t,1:n}-\bar{y})=\sum_{t\in w}|\bar{y}_{t,1:n}-\bar{y}|.$$

When each $C_{t,1:n}=\{t\}$, then the total value to be explained is $\sum_{t\in w}|f(\boldsymbol{x}_t)-\bar{y}|$. Then, for large $f(\boldsymbol{x}_t)$, we explain $f(\boldsymbol{x}_t)-\bar{y}$, while for small $f(\boldsymbol{x}_t)$ we explain $\bar{y}-f(\boldsymbol{x}_t)$, so that in either case we are explaining $|\bar{y}-f(\boldsymbol{x}_t)|$. When $f(\boldsymbol{x}_t)=\bar{y}$, then we are explaining an effect of 0. That may still involve offsetting positive

and negative effects, and not knowing a good sign to attrribute to them we count them as zero. The aggregate Shapley value of variable $j$ is then

$$\phi_j^w = \sum_{t \in w} \mathrm{sgn}_t \times \phi_{j,t} \tag{8}$$

by the additivity property.

This signed aggregation is not limited to cohort Shapley. It could also be applied to baseline Shapley. For baseline Shapley, we could aggregate over targets and/or baselines.

# 6 Examples

In this section we include some numerical examples of cohort Shapley. Section 6.1 computes CS for passengers for predicted probability of survival on the Titanic. It also computes some aggregate cohort Shapley values there. Section 6.2 computes CS for the Boston housing data and includes a comparison to baseline Shapley.

## 6.1 Titanic data

Here we consider a subset of the Titanic passenger dataset containing 887 individuals with complete records. This data has been used by Kaggle (see `https://www.kaggle.com/c/titanic/data`) to illustrate machine learning. As the function of interest, we construct a logistic regression model which predicts 'Survival' based on the predictors 'Pclass', 'Sex', 'Age', 'Siblings.Spouses.Aboard', 'Parents.Children.Aboard', and 'Fare'. Our model outputs an estimated probability of survival, $f(\boldsymbol{x}_t) \in [0, 1]$. To calculate the cohort Shapley values, we define similarity as exact for the discrete predictors 'Pclass', 'Sex', 'Siblings.Spouses.Aboard', and 'Parents.Children.Aboard' and a distance less than 1/20 of the variable range on the continuous predictors 'Age' and 'Fare'.

Figure 1 shows the cohort Shapley values for each predictor stacked vertically for every individual. The individuals are ordered by their predicted survival probability. Starting at zero, we plot a blue bar up or down according to the cohort Shapley value for the sex variable. Then comes a yellow bar for Pclass and so on.

A visual inspection of Figure 1 reveals clusters of individuals with similar Shapley values for which we could potentially develop a narrative. As just one example, we see passengers with indices between roughly 325 and 500 who have negative Shapley values for 'Sex' but positive Shapley values for 'Pclass' while their predicted value is below the mean. Many of the these passengers are men who are not in the lowest class.

We also report some aggregate cohort Shapley values in Table 6.1 given by equation (8). We see that 'Sex' has a substantially larger aggregate impact than the other predictors. We can further dig into subgroups to see how the impact varies with the covariates. For example, 'Sex' has a far greater impact

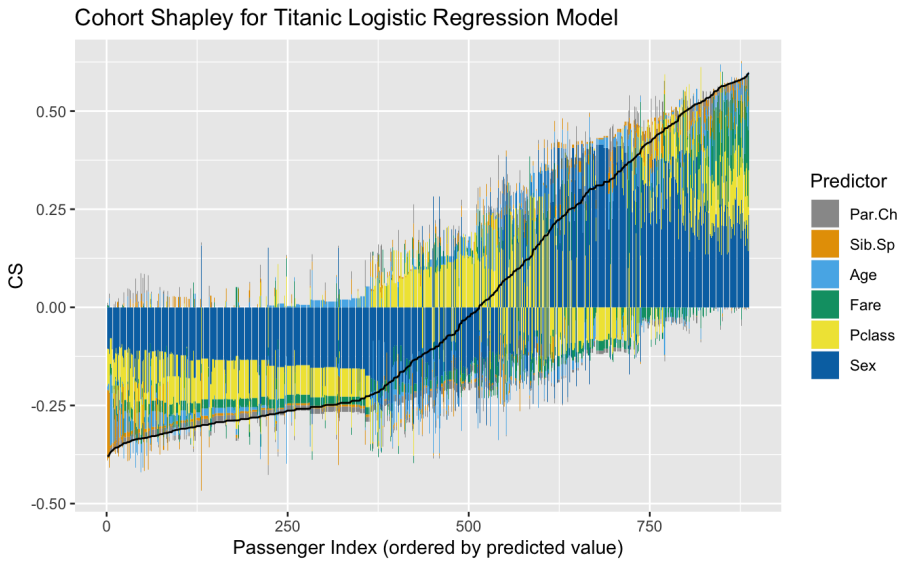


Figure 1: Cohort Shapley values stacked vertically for all passengers, ordered by estimated survival probability. The black overlay is $f(\boldsymbol{x}_t) - \bar{y}$ for each passenger.

for women where being female predicts a higher survival rate than for men where being male predicts a lower survival rate. Similarly we see a disparity in the aggregate impact for 'Pclass' between 1st and 3rd class women as both groups on average are more likely to survive than the average passenger, and being in 1st class contributes to that positive residual while being in 3rd class detracts.

In the final column of Table 6.1, we consider the same data, but instead of using fitted values from logistic regression as our function of interest, we use the actual survival outcomes as our black box. We calculate the cohort Shapley values using the same characterization of similarity to obtain comparable aggregate cohort Shapley values. Though the impact of 'Age' is slightly higher for this model, overall we see very similar values. To some degree this is to be expected when the model fits the training data well.

## 6.2 Boston housing dataset

The Boston housing dataset has 506 data points with 13 predictors and the median house value as a response (Harrison and Rubinfeld, 1978). Each data point corresponds to a vicinity in the Boston area. We fit a regression model to predict the house price from the predictors using XGBoost (Chen and Guestrin, 2016).

This dataset is of interest to us because it includes some striking examples of dependence in the predictors. For instance, the variables 'CRIM' (a measure of per capita crime) and 'ZN' (the proportion of lots zoned over 25,000 square

| Average Aggregate Cohort Shapley | | | | | | |
|---|---|---|---|---|---|---|
| | Logistic Regression Model | | | | | True Model |
| Pred. | All | M | F | F & 1st | F & 3rd | All |
| Pclass | 0.04 | 0.05 | 0.02 | 0.13 | −0.07 | 0.05 |
| Sex | 0.17 | 0.12 | 0.27 | 0.24 | 0.27 | 0.16 |
| Age | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.05 |
| Sib.Sp | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.03 |
| Par.Ch | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 |
| Fare | 0.02 | 0.02 | 0.02 | 0.11 | −0.02 | 0.03 |
| Avg $f(\boldsymbol{x}_t)$ | 0.39 | 0.19 | 0.74 | 0.92 | 0.60 | 0.39 |

Table 3: Aggregate cohort Shapley per person among various groups for the logistic regression model and the true value model. 1st and 3rd refer to values of 'Pclass' and the final row is the mean fitted value within the group.

feet) can be either near zero or large, but none of the 506 data points have both of them large and similar phenomena are in other scatterplots that we show below.

We will compute baseline Shapley and cohort Shapley for one target point. That one is the 205'th case in the sklearn python library and also in the mlbench R package. It is the unique one with 'RM'= 8.034. This target was chosen to be one for which some synthetic points in baseline Shapley would be far from any real data, but we did not optimize any criterion measuring that distance, and many of the other 506 points share that property. For cohort Shapley, we consider predictor values to be similar if their distance is less than 1/10 of the difference between the 95'th and 5'th percentiles of the predictor distribution.

Figure 2 shows two scatterplots of the Boston housing data. It marks the target and baseline points, depicts the cohort boundaries and it shows housing value in gray scale. The baseline point is $\boldsymbol{x}_b = (1/s)\sum_{i=1}^{s}\boldsymbol{x}_i$, the sample average, and it is not any individual subject's point partly because it averages some integer valued predictors. Here, the predicted house prices are 28.38 for the subject and 13.43 for the baseline. The figure also shows some of the synthetic points used by baseline Shapley. Some of those points are far from any real data points even in these two dimensional projections. There is a risk that the model fits for such points are not well determined.

Figure 3 shows baseline Shapley values for this target subject. We see that 'CRIM', 'RM', and 'LSTAT' have very large impact and the other variables do not. Figure 4 shows cohort Shapley values for this same subject. For cohort Shapley, the most impactful predictors are 'RM', 'ZN' and 'LSTAT' followed by a very gradual decline.

In baseline Shapley 'CRIM' was the most important variable, while in cohort Shapley it is one of the least important variables. We think that the explanation is from the way that baseline Shapley uses the data values at the upper orange cross in the top plot of Figure 2. The predicted price for a house at the synthetic

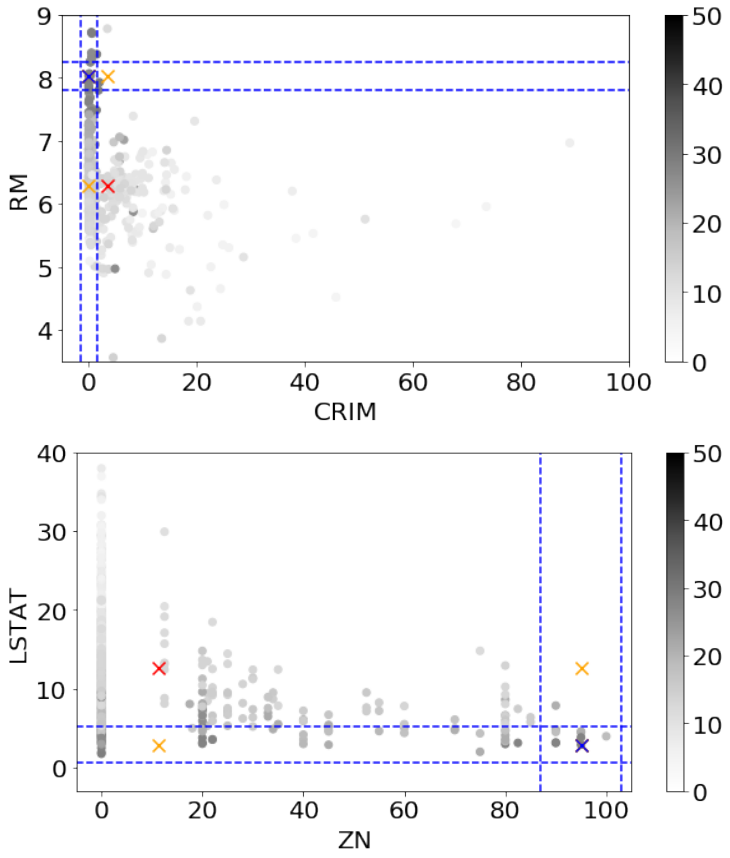


Figure 2: Two scatterplots of the Boston housing data. The target point is a blue X. The baseline is a red X. Synthetic points used by baseline Shapley are orange X's. Dashed blue lines delineate the cohorts we used.

point given by the upper orange cross is 14.17, which is much smaller than that of the subject, and even quite close to the baseline mean. This leads to the impact of 'CRIM' being very high. Data like that synthetic point were not present in the training set and so that value represents an extrapolation where we do not expect a good prediction. We believe that an unreliable prediction there gave the extreme baseline Shapley value that we see for 'CRIM'.

Related to the prior point, refining the cohort on 'RM' reduces its cardinality much more than refining the cohort on 'CRIM' does. Because cohort Shapley uses averages of actual subject values, refining the target on 'CRIM' removes fewer subjects and in this case makes a lesser change.

The lower panel in Figure 2 serves to illustrate the effect of dependent predictors on cohort Shapley value. The model for price hardly uses 'ZN', if at all, and the baseline Shapley value for it is zero. Baseline Shapley atributes a large

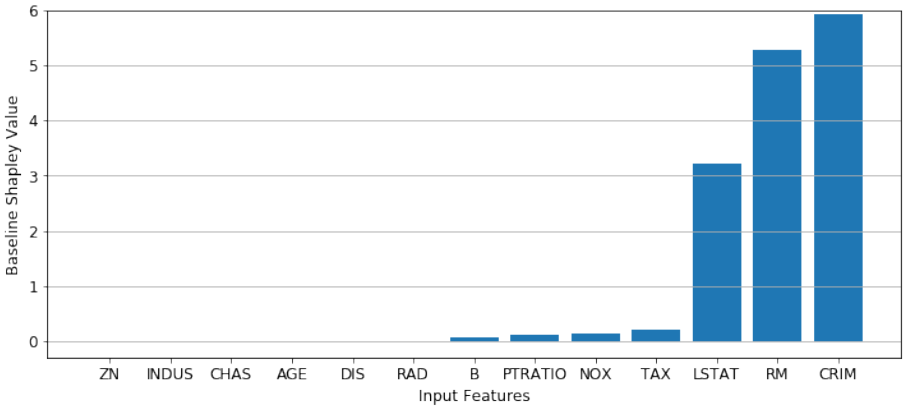


Figure 3: Baseline Shapley values for subject 205 of the Boston housing data.

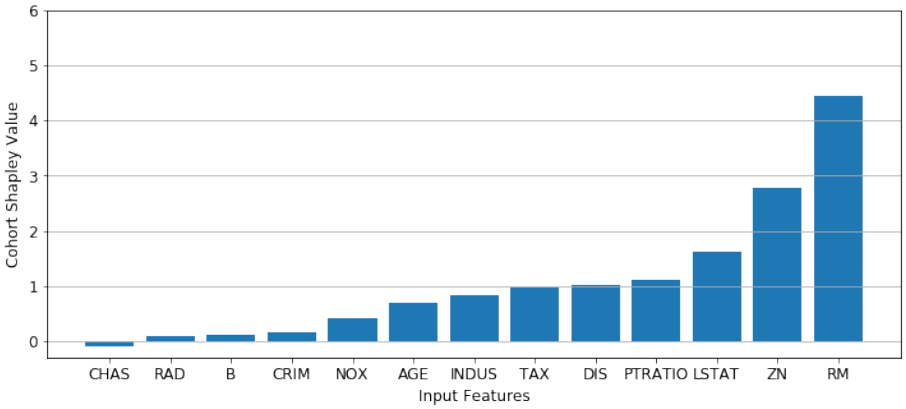


Figure 4: Cohort Shapley values for subject 205 of the Boston housing data.

impact to 'LSTAT' and nearly none to 'ZN'. For either of those predictors, the cohort mean is higher than the global average, and both 'LSTAT' and 'ZN' have high impact in cohort Shapley.

We can explain the difference as follows. As 'ZN' increases, the range of 'LSTAT' values narrows, primarily by the largest 'LSTAT' values decreasing as 'ZN' increases. Refining on 'ZN' has the side effect of lowering 'LSTAT'. Even if 'ZN' itself is not in the model, the cohort Shapley value captures this effect. Baseline Shapley cannot impute a nonzero impact for a variable that the model does not use. We say more about this in Section 7.

# 7 Discussion

Cohort Shapley resolves two conceptual problems in baseline Shapley and many other methods. First, it does not use any impossible or even unseen predictor combinations. Second, if two predictors are identical then it is perfectly predictable that their importances will be equal rather than subject to details of which black box model was chosen or what random seed if any was used in fitting that model.

Baseline Shapley and cohort Shapley have different counterfactuals and they address different problems. In baseline Shapley, we start with predictors at a baseline level and consider the effects of moving them one at a time towards the target point. In cohort Shapley, we start behind a 'veil of ignorance' knowing only that the subject was in the data set and then reveal information about the predictors one at a time to focus on subjects more like the target. Baseline Shapley helps us understand what a target subject might have done differently while cohort Shapley helps us understand which predictors influenced the comparison of the target subject to the other subjects.

If a predictor $x_j$ is never used by the black box $f$, then it will have a baseline Shapley value of zero. If that variable is correlated with a predictor that is used, then $x_j$ may still get nonzero impact in cohort Shapley. Knowing the value of $x_{tj}$ tells us something about $f(\boldsymbol{x}_t)$. For example, in some medical settings, we might find that the race of a patient has a cohort Shapley impact on their treatment even if their race was not used when $f$ was constructed.

Cohort Shapley requires $2^n$ quantities when there are $n$ predictor variables, and so for large $n$ it could be infeasible to compute it exactly. This is common to many but not all Shapley methods. For instance, Lundberg and Lee (2017) note that in tree structured algorithms many fewer than $2^n$ combinations may be required.

Cohort Shapley requires user input to define similarity. This is a strength and a weakness. It is a weakness because it places a burden on the user, while at the same time a strength in cases where the user has domain knowledge about what makes feature levels similar. There is a related literature on how finely a continuous variable should be broken into categories. Gelman and Park (2009) suggest as few as three levels for the related problem of choosing a discretization prior to fitting a model. We have used more levels but this remains an area for future research.

Cohort Shapley depends on the average predicted value in some potentially small cohorts, perhaps even the singleton for the target subject. Those predictions are normally made based on all $s$ observations and subject to regularization. As a result, $\bar{y}_{t,u}$ or even $f(\boldsymbol{x}_t)$ itself, does not necessarily have a large variance. If however, $\boldsymbol{x}_t$ is in an unusual part of the input space, then $\bar{y}_{t,u}$ for large $|u|$ might be poorly determined due either to bias or variance. In such cases, we might see unusual importance scores. Those scores retain their interpretation as explanations for why $f(\boldsymbol{x}_t)$ differs from the average prediction, and if they are clearly intuitively unreasonable, then they serve to reveal problems in $f$.

## Acknowledgments

We thank Masashi Egi of Hitachi for valuable comments. This work was supported by the U.S. National Science Foundation under grant IIS-1837931.

## References

Aliş, Ö. F. and Rabitz, H. (2001). Efficient implementation of high dimensional model representations. *Journal of Mathematical Chemistry*, 29(2):127–142.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Gamboa, F., Janon, A., Klein, T., Lagnoux, A., and Prieur, C. (2016). Statistical inference for Sobol' pick-freeze Monte Carlo method. *Statistics*, 50(4):881–902.

Gelman, A. and Park, D. K. (2009). Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician*, 63(1):1–8.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

Jiang, T. and Owen, A. B. (2003). Quasi-regression with shrinkage. *Mathematics and Computers in Simulation*, 62(3-6):231–241.

Kuo, F., Sloan, I., Wasilkowski, G., and Woźniakowski, H. (2010). On decompositions of multivariate functions. *Mathematics of computation*, 79(270):953–966.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

Molnar, C. (2018). *Interpretable machine learning: A Guide for Making Black Box Models Explainable*. Leanpub.

Owen, A. B. (2013). Variance components and generalized Sobol' indices. *Journal of Uncertainty Quantification*, 1(1):19–41.

Owen, A. B. (2014). Sobol' indices and Shapley value. *Journal on Uncertainty Quantification*, 2:245–251.

Owen, A. B. and Prieur, C. (2017). On Shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, New York. ACM.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global Sensitivity Analysis. The Primer*. John Wiley & Sons, Ltd, New York.

Shapley, L. S. (1952). A value for n-person games. Technical report, DTIC Document.

Sobol', I. M. (1969). *Multidimensional Quadrature Formulas and Haar Functions*. Nauka, Moscow. (In Russian).

Sobol', I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414.

Sobol', I. M. (2003). Theorems and examples on high dimensional model representation. *Reliability Engineering & System Safety*, 79(2):187–193.

Song, E., Nelson, B. L., and Staum, J. (2016). Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083.

Štrumbelj, E. and Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of machine learning research*, 11:1–18.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.

Sundararajan, M. and Najmi, A. (2019). The many Shapley values for model explanation. In *Proceedings of the ACM Conference '17*, New York. ACM.

Zhao, Q. and Hastie, T. (2019). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, pages 1–19. to appear.