

Supplementary Information (SI)

1.1 SHAP Properties

One property, namely *local accuracy* (known as *additivity* in game theory) was shown to be preserved in section 2.1; we now address the remaining two.

Consistency (known as *monotonicity* in game theory) states that if a model changes so that some feature’s contribution increases or stays the same regardless of the other inputs, that input’s attribution should not decrease:

Property 1 (Consistency / Monotonicity) *For any two models f and f' , if*

$$f'_x(S) - f'_x(S \setminus i) \geq f_x(S) - f_x(S \setminus i) \quad (7)$$

for all subsets of features $S \in \mathcal{F}$, then $\phi_i(f', x) \geq \phi_i(f, x)$.

Consistency extends to our smoothed SHAP scores for informative features ($\gamma, \hat{\beta}_j > 0$) since $\phi_i(f', x) \geq \phi_i(f, x) \Rightarrow \gamma \cdot \hat{\beta}_j \cdot \phi_i(f', x) \geq \gamma \cdot \hat{\beta}_j \cdot \phi_i(f, x)$. For the case of non-informative features which might produce small negative slopes, there could be a minor violation of consistency. However, we would argue that (i) this would affect only noisy SHAP scores which are likely of little interest anyhow and (ii) if consistency was critical, we could use the absolute values in our definition: $\hat{\phi}_j(f, x) \equiv \gamma \cdot |\hat{\beta}_j| \cdot \phi_j(f, x)$

Missingness (similar to *null effects* in game theory) requires features with no effect on the set function f_x to have no assigned impact.

Property 2 (Missingness) *If*

$$f_x(S \cup i) = f_x(S) \quad (8)$$

for all subsets of features $S \in \mathcal{F}$, then $\phi_i(f, x) = 0$.

Missingness extends trivially to our smoothed SHAP scores since $\phi_i(f, x) = 0 \Rightarrow \hat{\phi}_i(f, x) = \gamma \cdot \hat{\beta}_j \cdot \phi_i(f, x) = 0$

1.2 SAGE

We show evidence that SAGE values (Covert et al., 2020) are not immune to the cardinality bias as indicated in Figure 6. Though x_2 is the only informative feature (moderate SNR, $r = 0.15$), its importance is masked by x_1 , and to some degree by x_5 .

1.3 XGBoost results

In this section we provide further results for additional models and data considered. Figure 7 is the analogue of Figure 3 (same simulated data) with the following differences.

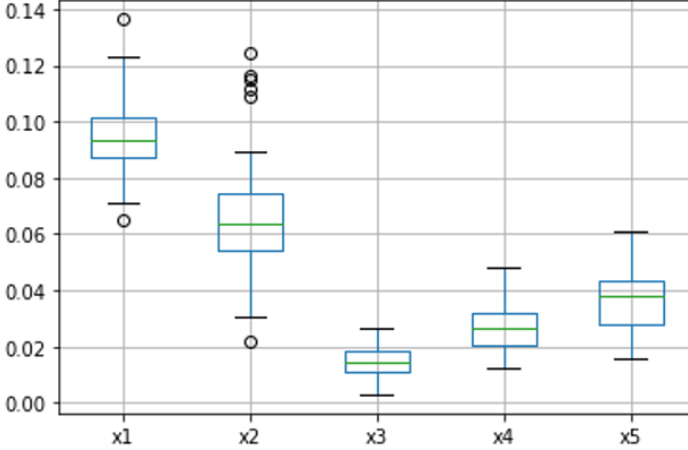


Fig. 6: *Shapley additive global importance* (SAGE) scores for the simulated data introduced in Figure 1 (power case, $r = 0.15$). The model is XGboost (sklearn, default parameters).

- Instead of random forests we used XGBoost⁶
- Boosting does not use the bootstrap for sampling, so we refrain from the *inbag/outbag* shrinkage.

We conclude from this example that our method is equally successful in debiasing SHAP scores for boosted trees as it is for random forests.

1.3.1 Complex data

As was already highlighted by Baudeau et al. (2022), the simulated classification data proposed by Strobl et al. (2007) has an extremely simple structure, in fact too simple to truly justify the use of complex models such as forests or boosted trees. We now investigate the additional regression function

$$Y = 10 \sin(\pi X_1 X_2) + 20(X_3 - 0.05)^2 + 10X_4 + 5X_5 + \epsilon \quad (9)$$

which Mentch and Zhou (2020) refer to as ‘MARS’. Unlike the previous generating process, it contains a nonlinearity as well as explicit interactions between the features. Features in the MARS model are sampled independently from $\text{Unif}(0, 1)$ and the noise as $\epsilon \sim N(0, \sigma^2 I)$. The variance σ^2 of the noise term is chosen so as to induce particular SNRs ν , which for the simulations below we consider 6 SNR values $\log_{10}\nu = -3, -2.3, -2, -1.7, -1.3, -0.8$ close to equally spaced on the log scale. Our first experiment adds 2 extra non-informative features $X_6, X_7 \sim \text{Unif}(0, 1)$ to Eq. (9), the results of which are shown in Figure 8.

⁶<https://cran.r-project.org/web/packages/xgboost/index.html> with `max_depth= 6` and `ntrees= 200`

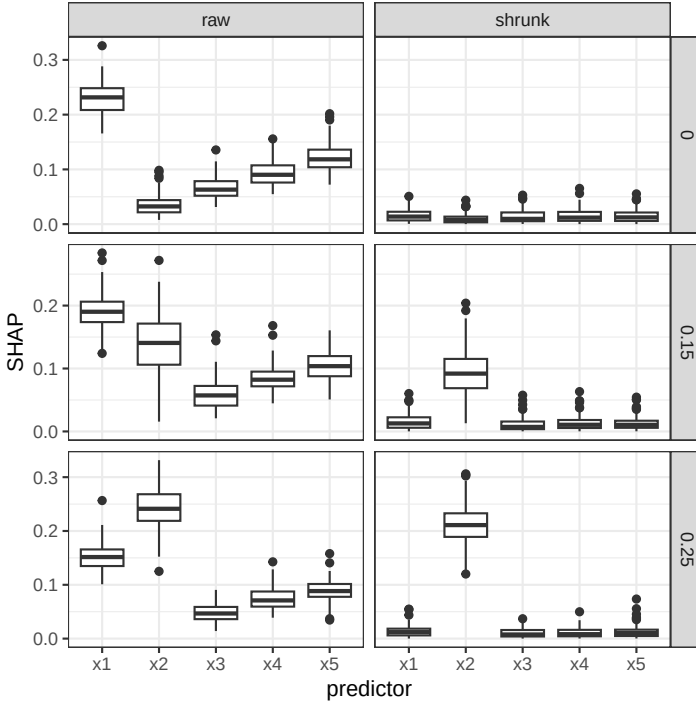


Fig. 7: Results for xgboost instead of RF for the simulated data described in section 1.1. Raw vs. shrunk SHAP scores for the null (top row) as well as 2 power simulations where only feature X_2 is informative ($r = 0.15, 0.25$ in middle and bottom row, respectively). No oob/inbag sample split is shown; the "shrunk" column refers to the train/test method.

While it is difficult to derive "ground truth" SHAP values for this dataset⁷, we can consider the highest SNR depicted in the lowest panel as an approximation to the true feature importances (since all features are continuous, there is no undue splitting preference). The upper panel -with its extremely low SNR- is akin to a Null simulation. The "raw" SHAP scores in the 3-4 upper panels display significant distortion, putting the noisy features on par with the rest and lack discerning power in the ranks as well. The smoothed SHAP scores fare much better: most feature importances are "shrunk" close to zero for low SNRs, while the 2-3 lower panels seem to recover the true underlying structure reasonably well.

Our second experiment changes the originally continuous variables X_1, X_2, X_3 to (balanced) binary features in order to investigate the splitting bias. In version 1, displayed in Figure 9, there are only 5 predictors in total,

⁷We are forcing additive decompositions on a structure with interactions. It would be more sensible -but beyond the scope of this paper- to compute interaction SHAP values.

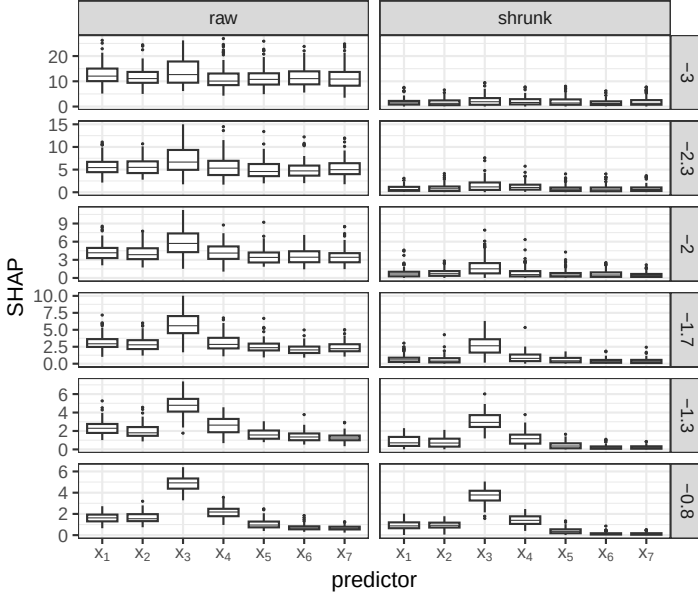


Fig. 8: Xgboost with parameters as in Figure 7 trained on the MARS data Eq. (9) augmented by non-informative features $X_6, X_7 \sim \text{Unif}(0, 1)$. The signal strength on the vertical panels is expressed by $\log_{10}(\text{SNR})$ which ranges from -3 to -0.8 . Note the lack of power to discern relevant features for low values of SNR. The smoothed SHAP scores for X_6, X_7 are correctly "shrunk" toward zero. And for high SNR, the ranking of and relative differences between the raw and smoothed scores are very similar.

whereas for version 2, displayed in Figure 10, we added again 2 extra non-informative features $X_6, X_7 \sim \text{Unif}(0, 1)$. In both figures, the bias toward the continuous features X_4, X_5 (and i.a. X_6, X_7) for the original SHAP scores is very pronounced, at least for SNRs below 0.05. The smoothed SHAP scores stay much closer to a reasonable ranking and in particular do not attribute any weight to X_6, X_7 .

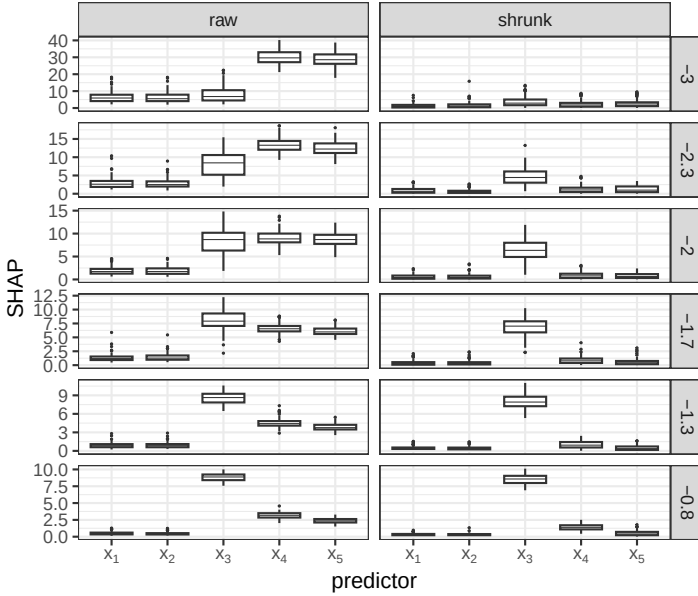


Fig. 9: Similar setting as in Figure 8 but (i) we did not add noisy features X_6, X_7 , and (ii) X_1, X_2, X_3 are now (balanced) binary features. The bias toward the continuous features X_4, X_5 in the left panels is very pronounced, at least for the first 5 rows. The smoothed SHAP scores seem to stay much closer to a reasonable ranking. On the other hand, the effective contribution of a binary variable is different than that from a continuous variable with the same range.

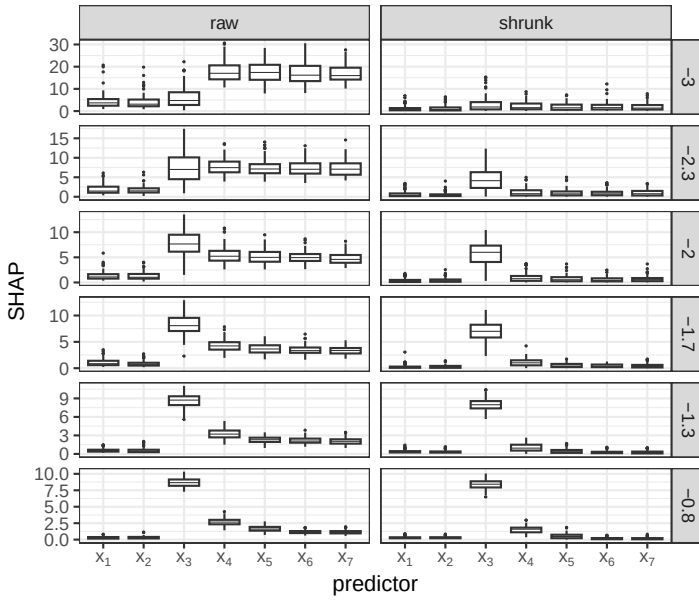


Fig. 10: Same setting as in Figure 9 but we added noisy features X_6, X_7 . The conclusion is quite similar