AStA Advances in Statistical Analysis
https://doi.org/10.1007/s10182-023-00479-7

ORIGINAL PAPER

Debiasing SHAP scores in random forests

Markus Loecher[1]

Received: 29 November 2022 / Accepted: 12 July 2023

Abstract
Black box machine learning models are currently being used for high-stakes decision making in various parts of society such as healthcare and criminal justice. While tree-based ensemble methods such as random forests typically outperform deep learning models on tabular data sets, their built-in variable importance algorithms are known to be strongly biased toward high-entropy features. It was recently shown that the increasingly popular SHAP (SHapley Additive exPlanations) values suffer from a similar bias. We propose debiased or "shrunk" SHAP scores based on sample splitting which additionally enable the detection of overfitting issues at the feature level.

Keywords Interpretable machine learning · Feature importance · Random forests · SHAP values · Explainable artificial intelligence

1 Introduction

The unprecedented success of machine learning (ML) algorithms in diverse fields such as medicine, finance, biology, marketing, public health, image classification and natural language processing is truly revolutionary. In these applications it is often important to have models that are both accurate and interpretable, where being interpretable means that we can understand how the model uses the input features to make predictions. Given a data input, how did an algorithm arrive at the output? Which inputs influenced the decision of the algorithm? More specifically, why did the model, e.g., reject the loan application of a person? Which risk factors led to, e.g., the prediction of a high risk of heart disease?

Many approaches have been proposed to address the lack of interpretability in ML. For an overview, see Molnar (2022). In bioinformatics, variable importance measures for random forests have gained popularity, particularly for selecting relevant genetic markers to predict diseases. Additionally, they have served as screening

✉ Markus Loecher
markus.loecher@hwr-berlin.de

[1] Department of Economics, Berlin School of Economics and Law, 10825 Berlin, Germany

Published online: 22 August 2023

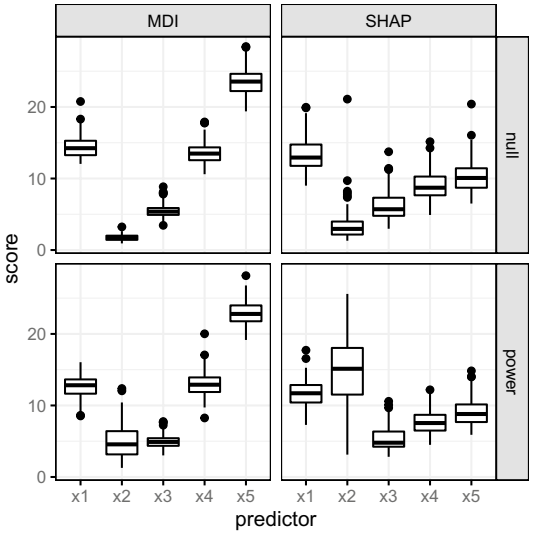

Fig. 1 Distributions of global feature importance scores (MDI and SHAP) for random classification forests using five features of different cardinalities (details of which are explained in the text). Upper row: Null simulation where no feature is informative; lower row: Power simulation where only feature X_2 affects the outcome ($r = 0.15$). Both simulations illustrate the severe bias exhibited by MDI and—to a lesser extent—SHAP scores which inflate the importances of high cardinality features $X_{1,3,4,5}$

tools in significant applications, emphasizing the importance of reliable and well-understood feature importance measures.

The overall goal of these approaches is to attribute importance to the features of a model according to their contribution to the model outcome. The importance of features in tabular data can be quantified at a local and global level. That is, how important is a feature for a particular decision (local) and how important is a feature for the overall decision-making process (global).

Tree-based ensemble methods such as random forests and gradient boosted trees achieve state-of-the-art performance in many domains and have consistently outperformed deep neural network models on tabular-style datasets so common in medicine and public health (Grinsztajn et al. 2022). The default choice to measure *global feature importances* in most software implementations of random forests is the *mean decrease in impurity (MDI)*. The MDI of a feature is computed as a (weighted) mean of the individual trees' improvement in the splitting criterion produced by each variable. A well-known shortcoming of this default measure is its evaluation on the in-bag samples which can lead to severe distortions of the reported ranking (Kim and Loh 2001; Strobl et al. 2007a). This bias is shown in the left panels of Fig. 1 for a simulation design used by Strobl et al. (2007a) with details as follows.

1.1 Data with varying cardinality

A binary response variable Y is predicted from a set of 5 predictor variables that vary in their scale of measurement and number of categories. The first predictor variable X_1 is continuous, while the other predictor variables $X_2, \ldots, X_5$ are multinomial with 2, 4, 10, 20 categories, respectively. The sample size for all simulation studies was set to n = 120. In the first *null case*, all predictor variables and the response are sampled independently. We would hope that a reasonable

variable importance measure would not prefer any one predictor variable over any other. In the second simulation study, the so-called *power case*, the distribution of the response is a binomial process with probabilities that depend on a parameter $r \in [0;0.5]$ which is related to the signal-to-noise-ratio (SNR) of x_2, namely $P(y = 1|X_2 = 1) = 0.5 - r, P(y = 1|X_2 = 2) = 0.5 + r$. For all figures, we repeatedly (100 times) fitted a random forest using the R[1]. ranger[2]. library with 400 classification trees and default parameters. SHAP values were computed using the R treeshap[3]. library.

1.2 Conditional feature contributions (CFCs)

The conventional wisdom of estimating the impact of a feature in tree-based models is to measure the *node-wise reduction of a loss function*, such as the variance of the output Y, and compute a weighted average of all nodes over all trees for that feature. By its definition, such a *mean decrease in impurity* (MDI) serves only as a global measure and is typically not used to explain a *per-observation, local impact*. Saabas et al. (2019) proposed the novel idea of explaining a prediction by following the decision path and attributing changes in the expected output of the model to each feature along the path:

Let f be a decision tree model, x the instance we are going to explain, $f(x)$ the output of the model for the current instance, and $f_x(S) \approx E[f(x) \mid x_S]$ the estimated expectation of the model output conditioned on the set S of feature values, then—following Lundberg et al. (2019)—we can define the *Saabas value*[4] for the i'th feature as

$$\phi_i^s(f, x) = \sum_{j \in D_x^i} f_x(A_j \cup i) - f_x(A_j), \tag{1}$$

where D_x^i is the set of nodes on the decision path from x that split on feature i, and A_j is the set of all features split on by ancestors of j. Equation (1) results in a set of feature attribution values that sum up to the difference between the expected output of the model and the output for the current prediction being explained. When explaining an ensemble model made up of a sum of many decision trees, the CFCs for the ensemble model are defined as the sum of the CFCs for each tree.

1.3 SHAP values

Lundberg et al. (2020) point out that CFCs are strongly biased to alter the impact of features based on their distance from the root of a tree. This causes CFC scores to be inconsistent, which means one can modify a model to make a feature clearly

[1] https://www.r-project.org/

[2] https://github.com/imbs-hl/ranger.

[3] https://github.com/ModelOriented/treeshap.

[4] Synonymous with *conditional feature contributions*.

more important, and yet the CFC attributed to that feature will decrease. The authors extend the original Shapley values from game theory to explain machine learning models based on *SHapley Additive exPlanation* (SHAP). For trees, these SHAP values can be computed efficiently in polynomial time. SHAP values can be seen as a solution to this problem, though often incurring potentially unnecessary computational cost as pointed out in Loecher (2022). As explained in Lundberg et al. (2019), Shapley values are computed by introducing each feature, one at time, into a conditional expectation function of the model's output, $f_x(S) \approx E[f(x) \mid x_S]$, and attributing the change produced at each step to the feature that was introduced; then averaging this process over all possible feature orderings. Shapley values represent the only possible method in the broad class of *additive feature attribution methods* that will simultaneously satisfy three important properties: *local accuracy*, *consistency*, and *missingness*. They are defined as:

$$\phi_i(f,x) = \sum_{R\in\mathcal{R}} \frac{1}{M!}\left[f_x(P_i^R \cup i) - f_x(P_i^R)\right], \tag{2}$$

where $\mathcal{R}$ is the set of all feature orderings, P_i^R is the set of all features that come before feature i in ordering R, and M is the number of input features for the model. Local accuracy states that for a specific input x, the explanation's attribution values ϕ_i for each feature i need to sum up to the output $f(x)$:

$$f(x) = E(f) + \sum_{i=1}^{M} \phi_i(f,x) \tag{3}$$

In this paper we are mainly concerned with global importance measures, which is naturally defined as the average of the absolute SHAP values per feature across the data:

$$I_j = \frac{1}{N}\sum_{i=1}^{N} |\phi_j(f,x_i)| \tag{4}$$

All of our results pertain to marginal SHAP values.

1.4 Related work

Marginal Shapley values are used to calculate the average contribution of a feature to the prediction when all possible coalitions of features are considered. This approach assumes that each feature is independent of the others, which may not be the case in reality. When features are highly correlated, the marginal Shapley value for one feature may be misleading as it doesn't consider the influence of the other correlated features. On the other hand, *conditional* Shapley values consider the contribution of a feature based on a specific set of conditions. In other words, it takes into account the correlation among the features and calculates the average contribution of the feature based on the specific values of the correlated features. This approach provides a more discerning understanding of the contribution of each feature to the prediction.

Janzing et al. (2020) offer a causal interpretation of Shapley values that replaces the conventional conditioning by observation with conditioning by intervention.

Sundararajan and Najmi (2020) highlighted several drawbacks of the SHAP method, including the production of counterintuitive explanations when some features are deemed unimportant. However, the Baseline Shapley (BShap) method offers an improvement to this "uniqueness" issue in attribution methods.

At first glance similar to our findings, Kwon and Zou (2022) also demonstrate that Shapley values often fail to sort features in order of influence on a model prediction. These attribution mistakes are especially pronounced when different marginal contributions have different signal and noise. The authors also propose a reweighing scheme as a solution which they refer to as *WeightedSHAP*. However, the underlying reasons for the incorrect ordering of features are very different from our rather narrow situation which combines the greedy tree splitting algorithm with varying feature entropies.

Covert et al. (2020) introduce Shapley additive global importance (SAGE) for quantifying a model's dependence on each feature. These global measures are the result of minimizing a structured loss function and hence much less heuristic than the ones defined by Eq. (4). Sutera et al. (2021) demonstrated that SAGE applied to tree-based models is very similar to the Gini importance. Supplement 1.2 contains limited results on SAGE values for our simulated data introduced in Fig. 1. We refrained from delving deeper into this alternative method because (i) the provided code on the repository runs extremely slowly, and (ii) our focus are the original SHAP values which have become extremely popular.

Additional approaches have been suggested to compute global SHAP values (Frye et al. 2020; Williamson et al. 2020; Casalicchio et al. 2019) where the latter reference proposes local and global feature importance methods based on the ideas of partial dependence and ICE which are lend themselves particularly well for visualization. Nevertheless, it is important to note that the inherent bias resulting from the underlying tree structure persists across all of these methods, provided that the tree-building algorithm remains unchanged.

Yasodhara et al. (2021) conducted extensive experiments to assess the accuracy and stability of global feature importance scores, including SHAP and Gini. Their findings revealed notably low correlations with the true feature rankings, even in scenarios without additional noise. Moreover, when inputs or models were perturbed, these correlations further decreased. However, the authors did not specifically examine the impact of uninformative variables with different levels of cardinality.

2 Smoothed SHAP

Less known than the distortions w.r.t MDI is the persistence of the bias of feature importance rankings based on (global) SHAP values, as shown in the right panels of Fig. 1. Similar to MDI, SHAP shows a strong bias toward high-entropy variables which in the power case yields distorted rankings of variable impact. The grave consequences of providing such flawed explanations for, e.g., financial decisions or predictions of medical outcomes cannot be taken serious enough. To the best of

our knowledge, no debiasing method has been successfully applied to SHAP scores (yet).

2.1 Sample splitting

It was recognized early on that separating the split variable selection from the split point selection can reduce or eliminate the bias (Loh and Shih 1997; Hothorn et al. 2006). A somewhat similar approach is implemented in so-called "honest trees" (Athey and Imbens 2016). Honest trees are based on sample-splitting: Each node value is estimated using a different sub-sample of the training data than the one used to decide the split point. The sample-splitting successfully eliminates the split bias but comes with a cost of lower performance, as the number of samples for each estimation (split point and prediction) is reduced.

It is somewhat surprising that the clear differentiation between validation and training data, which is typically observed for most other metrics in statistical modeling, has not been consistently applied to variable importance measures. Recently, several authors (Li et al. 2019; Loecher 2020b; Zhou and Hooker 2021) effectively eliminated the outlined bias inherent to the tree splitting procedure by including out-of-train samples in order to compute a debiased version of the MDI importance. These *post-hoc* adjustments of scores derived from tree-based models do not alter the tree structure, as opposed to attempts to prevent overfitting directly during the growth of the tree, as described in, e.g., Adler and Painsky (2022).

The main idea is based on sample splitting in order to counteract the severe overfitting of high-cardinality features. Instead of evaluating feature importance scores solely on the same *in-bag* (IB) samples that were used to grow the trees, we penalize discrepancies with the same scores evaluated on *out-of-bag* (OOB) samples. While random forests—at least when fitted with the bootstrap—conveniently provide these IB/OOB without any "extra costs," we choose to generalize our presentation by splitting the data into train/test sets instead. This leads to a more model agnostic debiasing approach and hopefully simplifies the clarity of presentation. Whenever applicable, we show results for both train/test splits as well as IB/OOB splits. The initial steps of our proposed algorithm are as follows.

1. Split data into train/test sets. (Note the unusual terminology, since we are also "training" a model on "test" data)
2. Fit two separate random forest models on train/test each: RF_{train}, RF_{test}
3. Compute separate **in/out** SHAP values for, e.g., the test data

   (a) SHAP_{in}: RF_{test} predicts on test
   (b) SHAP_{out}: RF_{train} predicts on test

We would expect weak/no correlations between SHAP_{in} and SHAP_{out} for non-informative features and in turn strong correlations for relevant features, which is confirmed by Fig. 2.

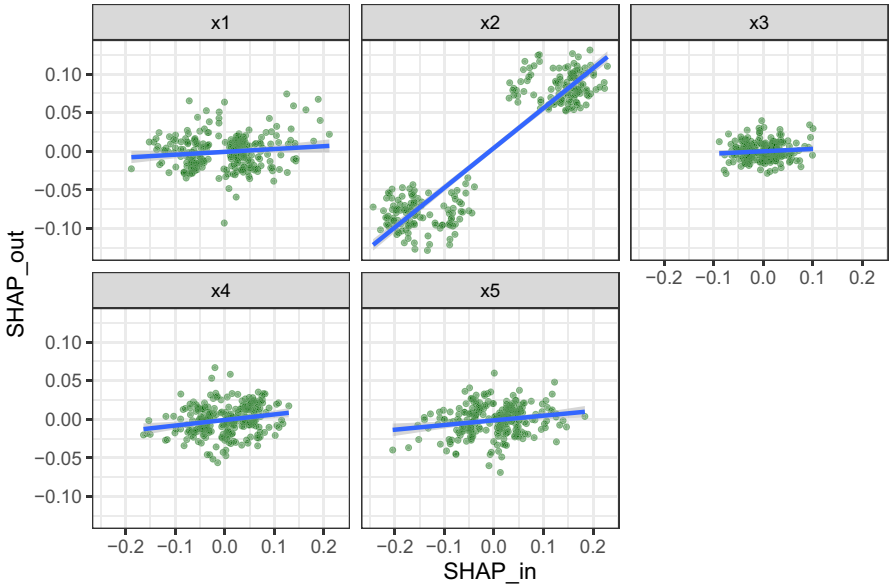

Fig. 2 In-sample versus out-sample SHAP scores for the power simulation where only feature X_2 is informative ($r = 0.25$), illustrating the low correlations for features $X_{1,3,4,5}$ and a strong relationship for X_2, with values of $R^2_{1,2,3,4,5} = [0.038, 0.91, 0.004, 0.05, 0.044]$ respectively. (sample size $n = 240$)

Based upon these two sets of in-sample and out-sample SHAP scores, we propose the following debiasing/smoothing/shrinkage scheme. For each feature j:

1 Fit a linear model $\text{SHAP}_{j,out} = \beta_j \cdot SHAP_{j,in} + u$

2 Use the estimates $\widehat{SHAP}^{shrunk}_{j,in} = \hat{\beta}_j \cdot SHAP_{j,in}$ as local explanations instead of $SHAP_{j,in}$,

We can think of these predictions as "smoothed" version of the original SHAP values.

We can avoid the need to fit two random forests by computing SHAP scores separately for each tree which was fitted on a specific bootstrapped subset (inbag). If the number of trees is sufficiently high, the union of these tree-wise disjoint inbag/oob data sets will "generously" cover the entire original data, i.e., each row would have been part of the oob set sufficiently often. By averaging over all trees we hence obtain the equivalent of RF_{train} and RF_{test} SHAP scores with just one random forest.

We expect the correlation between these inbag and oob SHAP values to be high for informative and nearly zero for uninformative features, which motivates the following equivalent but simpler smoothing method. For each feature j

1. Fit a linear model $SHAP_{j,oob} = \beta_j \cdot SHAP_{j,inbag} + u$

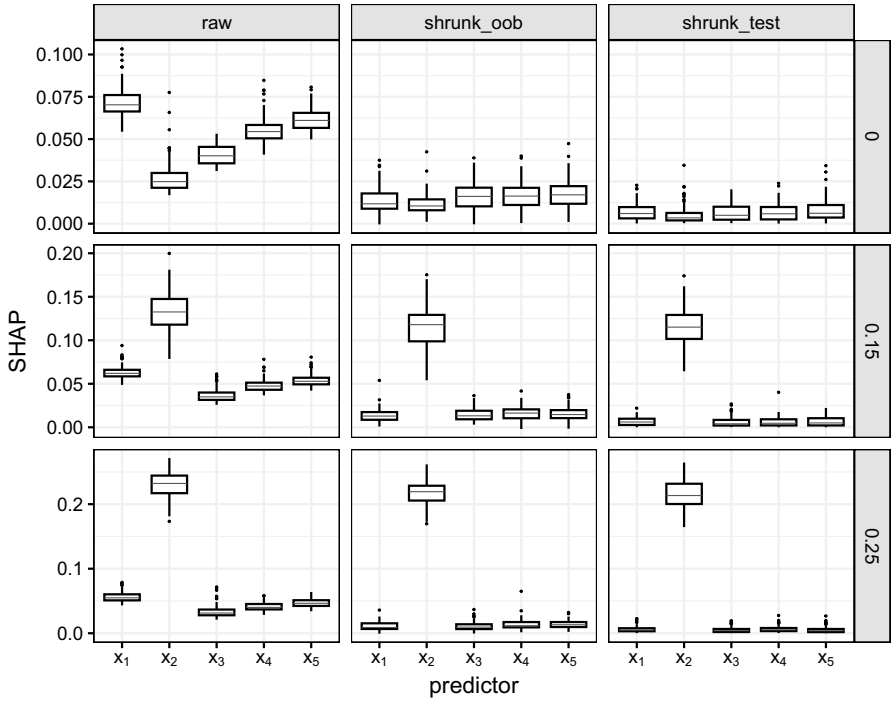

Fig. 3 Raw vs. shrunk SHAP scores for the null (top row) as well as 2 power simulations where only feature X_2 is informative ($r = 0.15, 0.25$ in middle and bottom row, respectively). The oob/inbag sample split (middle column) seems to lead to a similar debiasing effect as the computationally more expensive train/test method (right column)

2. Use the estimates $\widehat{SHAP}_{j,oob} = \hat{\beta}_j \cdot SHAP_{j,inbag}$ as local explanations instead of the original $SHAP_j$ which mix inbag and oob values.

The two right panels of Fig. 3 display the main results: the non-informative features are shrunk toward zero while the relevant variable x_2 is hardly modified. The benefits are most striking for low signal-to-noise ratio, e.g., the null simulation (top row) and for a moderate value of $r = 0.15$. We notice in passing that Kwon and Zou (2023) takes the idea of using oob data much further and propose them as a general data valuation method.

We emphasize that our main goal so far has been to debias global SHAP scores rather than local explanations; hence the emphasis on relative rankings and the loss of an absolute interpretable scale. If local accuracy as expressed by Eq. (3) shall be preserved, we can simply rescale the scores by multiplying with

$$\gamma = \sum_{j=1}^{M} \phi_j(f,x) / \sum_{j=1}^{M} \hat{\beta}_j \cdot \phi_j(f,x), \tag{5}$$

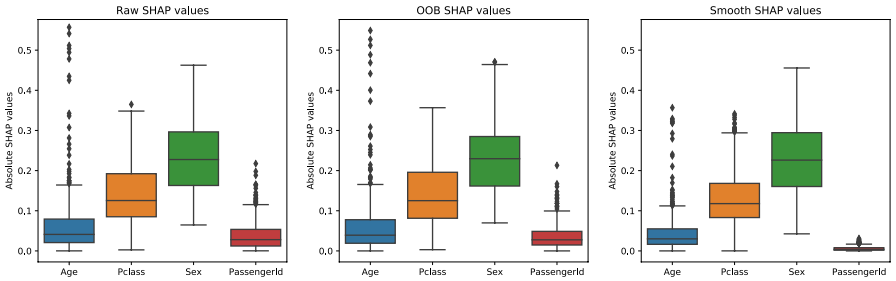

Fig. 4 Local SHapley Additive exPlanation (SHAP) values for the Titanic data. Original, "raw" (left panel) vs. SHAP scores computed only on oob data (middle panel) vs. shrunk, "smoothed" scores (right panel) as explained in the text

Fig. 5 Decomposition of a specific survival prediction for passenger with $ID = 42$ from the Titanic data. The original SHAP scores (upper panel) are highly variable and yield a large weight for *passengerID* while the smoothed explanations in the lower panel are (i) much less extreme, (ii) yield the same final prediction thanks to the applied rescaling in Eq. (5) and (iii) rank *passengerID* as a minute contribution

where $\phi_j(f,x) \equiv SHAP_{j,in}$ would be the "raw", unmodified SHAP score for feature j and input x. We redefine the smoothed scores as $\widehat{\phi}_j(f,x) \equiv \gamma \cdot \hat{\beta}_j \cdot \phi_j(f,x)$ which ensures that Eq. (3) still holds:

$$\begin{aligned} E(f) + \sum_{i=1}^{M} \widehat{\phi}_i(f,x) &= E(f) + \sum_{i=1}^{M} \gamma \cdot \hat{\beta}_j \cdot \phi_i(f,x) \\ &= E(f) + \sum_{j=1}^{M} \phi_j(f,x) \frac{\sum_{i=1}^{M} \hat{\beta}_j \cdot \phi_i(f,x)}{\sum_{j=1}^{M} \hat{\beta}_j \cdot \phi_j(f,x)} \\ &= E(f) + \sum_{j=1}^{M} \phi_j(f,x) \stackrel{!}{=} f(x) \end{aligned} \tag{6}$$

Fig. 5 serves to illustrate Eq. (6) by a specific example. Supplement 1.1 shows that the other two essential properties *consistency and missingness* are also unchanged.

2.2 Real data

Previous studies (Loecher 2020, 2022a) have used the well-known titanic data set to illustrate the severity of the bias of commonly used feature importance scores in random forests. In this section we show that even (SHAP) values suffer from (i) a strong dependence on feature cardinality, and (ii) assign non zero importance scores to uninformative features.

In the following model[5] we include *passengerID* as a feature along with the more reasonable *Age*, *Sex* and *Pclass*. The left panel of Fig. 4 shows the high variance of SHAP scores for *passengerID* which would hence result in a large global importance score. Simply separating the inbag from the oob SHAP values helps but is not a remedy as shown in the middle graph. Our proposed smoothing algorithm selectively shrinks the scores for *passengerID* and *Age* as seen in the right panel. While we strictly speaking have no "ground truth", we can make a very strong argument for *passengerID* to be a truly non-informative feature since the left panel boxplot barely changes when we randomly scramble the order of *passengerID*.

To close this section we examine the explanation of a specific prediction as decomposed by the original SHAP values, illustrated in the upper panel of Fig. 5 (referred to as a "force plot" in Lundberg et al. 2020). It seems highly implausible that the ID of the passenger would carry the second largest impact on the survival probability, while passenger class takes on a distant fourth rank. The smoothed explanations in the lower panel appear much more sensible; we also notice the reduced variability of the shrunk scores (The axes are not aligned).

2.3 Feature ranking as a classification task

For a more systematic study of the proposed shrunk SHAP scores, we closely follow the simulations outlined in Li et al. (2019) and pose the discrimination of relevant from noisy features as a classification task. Our generated data (2000 rows) contain 50 discrete features, with the jth feature taking on $j+1$ distinct values $0, 1, \ldots, j$. As shown in previous sections, discrete features with different number of distinct values constitute a critical challenge for MDI and SHAP. We randomly select a set S of 5 features from the first ten as relevant features. The remaining features are non-informative. All features are independent and—of course—samples are i.i.d. The binary dependent variable is generated using the following rule:

$$P(Y = 1|X) = \text{Logistic}\left(\frac{2}{5}\sum_{j\in S} x_j/j - 1\right)$$

Treating the noisy features as label 0 and the relevant features as label 1, we can evaluate a feature importance measure in terms of its area under the receiver operating characteristic curve (AUC). We grow 100 deep trees (minimum leaf size

[5] In all random forest simulations, we choose $mtry = 2, ntrees = 100$ and exclude rows with missing *Age*.

Table 1 Average AUC scores for relevant feature identification. The $\widehat{\text{SHAP}}^{shrunk}_{in}$ scores highlighted in bold face outperform all other methods

SHAP	$\widehat{\text{SHAP}}^{shrunk}_{in}$	SHAP_{oob}	MDA	MDI
0.66	**0.89**	0.73	0.65	0.10

MDA permutation importance, *MDI* (default) Gini impurity, SHAP_{oob} scores are based upon only the oob data. The $\widehat{\text{SHAP}}^{shrunk}_{in}$ scores outperform all other methods

equals 1, $m_{try} = 3$), repeat the whole process 100 times and report the average AUC scores for each method in Table 1. For completeness, we also compute the AUC for the permutation importance *MDA* which was long considered a gold standard but is not without its own issues (Hooker and Mentch 2019). For this simulated setting, $\widehat{\text{SHAP}}^{shrunk}_{in}$ achieves the best AUC score under all cases, most likely because it shrinks noise features toward zero. We notice that the AUC score for the OOB-only SHAP_{oob} outperforms the permutation importance and—unsurprisingly—the overall SHAP scores.

3 Discussion

We can make the proposed train/test split more efficient in various ways. The obvious one is exploiting the symmetry by also computing SHAP values for RF_{test} and RF_{train} on the train data:

1. SHAP_{in}: RF_{test} predicts on test, RF_{train} predicts on train
2. SHAP_{out}: RF_{train} predicts on test, RF_{test} predicts on train

One benefit would be the possibility to average the respective slopes, which should lead to more robust estimates. Alternatively we can view the above scheme as essentially akin to a 2-fold cross-validation with the obvious advantage that with enough computational resources one can generalize to a larger number of k-fold CV.

4 Conclusion

In this paper we have demonstrated that simple sample splitting can be utilized to substantially reduce the tendency of tree-based methods to overfit categorical features with large number of categories. The proposed debiasing scheme leans on the well-known statistical principle to not validate a model on the same ("in") data that were used to fit the model but instead to compute a goodness-of-fit measure on a separate ("out") data set. While perhaps most relevant to tree-based algorithms, our method is model agnostic and in fact its benefits extend beyond just debiasing SHAP values. We believe that the strength of the correlations between SHAP_{in} and SHAP_{out} scores could be used to **detect overfitting at a feature level** in the following sense.

While there are established tools (such as cross validation) to control model complexity in order to optimize performance on unseen data,

- Feature selection remains notoriously difficult
- The minima of, e.g., out-of-sample loss measures are typically rather broad allowing for a wide range of model parameters and feature subsets
- The consequences of choosing sub-optimal, overly adapted models (e.g., in terms of including non-informative features or allowing high local flexibility) are often
  - Rather modest/minor for predictive accuracy
  - But extremely misleading for model explanations
- This *"interpretational overfitting"* can be diagnosed at the feature level by inspecting the correlations between *in* and *out* SHAP values.

Summarizing, we view the well-established bias in Gini and SHAP importance measures as a specific case of overfitting and propose a *post-hoc*, easy-to-implement solution to this problem. We have illustrated its effectiveness on three simulated data sets as well as the well-known Titanic data.

It is crucial to acknowledge that the bias examined in this study is only considered a bias when we employ variable importance measures to infer the significance of variables in relation to underlying relationships within the data. However, if our interest lies in comprehending the tree-based model itself, then this "bias" actually represents a true effect. This is because the tree-based structure does utilize variables with more categories more frequently and places them closer to the top of the trees compared to variables with fewer categories. This distinction between biases is significant: a variable importance measure can exhibit bias in identifying which variable is important in determining the true outcome, as well as bias in identifying which variable is important in determining the predicted outcome, i.e., the model output. If the model being explained is not a perfect representation of the underlying relationship in the data, then any variable importance measure will demonstrate one of these biases. Hence, when employing a variable importance measure, it is crucial to always consider what the measure intends to explain: the model output or the real-world outcome.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10182-023-00479-7.

Acknowledgements ML thanks both the Berlin School of Economics and Law for non-financial research support and Philipp Heitmann and Raphael Franke for coding support as well as helpful discussions. ML also thanks the anonymous reviewer for very valuable and insightful comments which greatly improved this paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Data and code to reproduce all the results presented in this paper is available at https://github.com/markusloecher/shrunk_SHAP.

Declarations

Conflict of interest ML declares no conflict of interest.

References

Adler, A.I., Painsky, A.: Feature importance in gradient boosting trees with cross-validation feature selection. Entropy **24**(5), 687 (2022)

Athey, S., Imbens, G.: Recursive partitioning for heterogeneous causal effects. Proc. Natl. Acad. Sci. **113**(27), 7353–7360 (2016)

Baudeu, R., Wright, M., Loecher, M.: Are SHAP values biased towards high-entropy features?. In: ECML PKDD Workshop on XKDD. Springer (2022)

Casalicchio, G., Molnar, C., Bischl, B.: Visualizing the feature importance for black box models. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, pp. 655–670. Springer (2019)

Covert, I., Lundberg, S.M., Lee, S.-I.: Understanding global feature contributions with additive importance measures. Adv. Neural. Inf. Process. Syst. **33**, 17212–17223 (2020)

Frye, C., Rowat, C., Feige, I.: Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. Adv. Neural. Inf. Process. Syst. **33**, 1229–1239 (2020)

Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on tabular data?. arXiv preprint arXiv:2207.08815 (2022).

Hooker, G., Mentch, L.: Please stop permuting features: an explanation and alternatives. arXiv e-prints pp arXiv–1905 (2019)

Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. **15**(3), 651–674 (2006)

Janzing, D., Minorics, L., Blöbaum, P.: Feature relevance quantification in explainable AI: a causal problem. In: International Conference on artificial intelligence and statistics, pp. 2907–2916. PMLR (2020)

Kim, H., Loh, W.-Y.: Classification trees with unbiased multiway splits. J. Am. Stat. Assoc. **96**(454), 589–604 (2001)

Kwon, Y., Zou, J.: Data-oob: out-of-bag estimate as a simple and efficient data value. arXiv preprint arXiv:2304.07718 [cs.LG] (2023)

Kwon, Y., Zou, J.Y.: Weightedshap: analyzing and improving Shapley based feature attributions. Adv. Neural. Inf. Process. Syst. **35**, 34363–34376 (2022)

Li, X., Wang, Y., Basu, S., Kumbier, K., Yu, B.: A debiased MDI feature importance measure for random forests. In: Wallach, H., Larochelle, H., Beygelzimer A., d Alché-Buc, F., Fox E., Garnett, R. (eds) Advances in Neural Information Processing Systems, vol. 32, pp 8049–8059 (2019)

Loecher, M.: Unbiased variable importance for random forests. Commun. Stat. Theory Methods (2020). https://doi.org/10.1080/03610926.2020.1764042

Loecher, M.: Debiasing MDI Feature Importance and SHAP Values in Tree Ensembles. In: International cross-domain conference for machine learning and knowledge extraction, pp 114–129. Springer (2022a)

Loecher, M., Lai, D., Wu, Q.: Approximation of SHAP values for randomized tree ensembles. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, pp 19–30. Springer (2022b)

Loh, W.-Y., Shih, Y.-S.: Split selection methods for classification trees. Stat. Sin. 815–840 (1997)
Lundberg, S.M., Erion, G.G., Lee, S.-I.: Consistent individualized feature attribution for tree ensembles (2019)
Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. **2**(1), 56–67 (2020)
Mentch, L., Zhou, S.: Randomization as regularization: a degrees of freedom explanation for random forest success. J. Mach. Learn. Res. **21**(1), 6918–6953 (2020)
Molnar, C.: Interpretable Machine Learning, 2nd edn. (2022). https://christophm.github.io/interpretable-ml-book
Saabas, A.: Treeinterpreter library (2019). https://github.com/andosa/treeinterpreter
Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform. (2007). https://doi.org/10.1186/1471-2105-8-25
Sundararajan, M., Najmi, A.: The many Shapley values for model explanation. In: International conference on Machine Learning, pp 9269–9278. PMLR (2020)
Sutera, A., Louppe, G., Huynh-Thu, V.A., Wehenkel, L., Geurts, P.: From global to local mdi variable importances for random forests and when they are Shapley values. Adv. Neural. Inf. Process. Syst. **34**, 3533–3543 (2021)
Williamson, B., Feng, J.: Efficient nonparametric statistical inference on population feature importance using Shapley values. In: International Conference on Machine Learning, pp 10282–10291. PMLR (2020)
Yasodhara, A., Asgarian, A., Huang, D., Sobhani, P.: On the trustworthiness of tree ensemble explainability methods. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, pp 293–308. Springer (2021)
Zhou, Z., Hooker, G.: Unbiased measurement of feature importance in tree-based methods. ACM Trans. Knowl. Discov. Data (TKDD) **15**(2), 1–21 (2021)