Are SHAP values biased towards high-entropy features?

Raphael Baudeu[1][0000−0001−6617−7894], Marvin Wright[2,3,4][0000−0002−8542−6291], and Markus Loecher[5][0000−0002−6823−1994]

[1] University of Glasgow, Glasgow, UK
[2] Leibniz Institute for Prevention Research & Epidemiology – BIPS, Bremen, Germany
[3] Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany
[4] Department of Public Health, University of Copenhagen, Copenhagen, Denmark
[5] Berlin School of Economics and Law, Berlin, Germany

Abstract. In this paper, we examine the bias towards high-entropy features exhibited by SHAP values on tree-based structures such as classification and regression trees, random forests or gradient boosted trees. Previous work has shown that many feature importance measures for tree-based models assign higher values to high-entropy features, i.e. with high cardinality or balanced categories, and that this bias also applies to SHAP values. However, it is unclear if this bias is a major problem in practise or merely a statistical artifact with little impact on real data analyses. In this paper, we show that the severity of the bias strongly depends on the signal to noise ratio (SNR) in the dataset and on adequate hyperparameter tuning. In high-SNR settings, the bias is still present but is unlikely to affect feature rankings and thus can be safely ignored in many real data applications. On the other hand, in low-SNR settings, a feature without ground-truth effect but with high entropy could be ranked higher than a feature with ground-truth effect but low entropy. Here, we show that careful hyperparameter tuning can remove the bias.

1 Introduction

With the rising popularity of machine learning also came a growing need to understand prediction models through interpretability methods. These methods can have differing aims ranging from visualising the workings of the black-box model to feature importance measures to understand which features are the most important to the model or to some underlying relationship. One such interpretability method is SHAP (SHapley Additive exPlanations) [1]. SHAP values are based on Shapley values [2], a method from game theory that was first introduced to machine learning applications by Štrumbelj & Kononenko [3]. Lundberg & Lee [1] later introduced SHAP as the only additive feature attribution method satisfying the properties of local accuracy, missingness and consistency. Lundberg et al. [4] introduced a new method for efficiently calculating SHAP values for tree based structures (TreeSHAP).

Several feature importance methods for tree-based methods have been shown to have a bias towards high-entropy features, i.e., numerical features with many unique values, categorical features with high cardinality or balanced category frequencies. This phenomenon was already shown by Breiman et al. [5] for classification and regression trees and later made popular for model-specific random forest feature importance [6–9]. Loecher [10, 11] and Adler & Painsky [12] have shown that the same kind of bias is rooted deeper in the underlying tree structure and is thus not exclusive to random forests and its model-specific feature importance measures but rather extends to SHAP values for random forests and also to other tree-based prediction methods such as gradient boosted trees.

In this paper, we show that the bias observed for SHAP values is only a problem when the signal to noise ratio (SNR) in the dataset is very low. The bias becomes less important the higher the SNR and, while it is still present, hardly matters in practise in settings with high SNR. We do this by performing simulation studies on datasets with different levels of SNR. We also use two real-world datasets with different levels of SNR to show the impact of SNR on the bias. We further show that proper hyperparameter tuning can remove the bias in SHAP values.

2 Simulation Study

To investigate whether SHAP values exhibit a bias towards high-entropy features for tree-based models we conducted a simulation study. We used a data generating process (DGP), which consists in first generating four categorical explanatory variables (features) X_1, X_2, X_3, X_4 with 2,4,10 and 20 categories [6] and one continuous explanatory variable $X_5 \sim \mathcal{N}(0, 1)$, respectively. Let $\boldsymbol{X}$ be the $n \times 5$ matrix of explanatory variables. We used a sample size of n = 100 throughout Section 2. All categorical variables were multinomially distributed with equal probabilities for each category. The categories were $1, .., n_c$ where n_c corresponds to the number of categories of the variable in question. The explanatory variables were simulated independently of each other. We used an outcome variable $\boldsymbol{O} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \epsilon$, where $\boldsymbol{\beta}$ is a 1×5 vector and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with variance σ^2 selected such that we attain a pre-specified signal to noise ratio (SNR). We define the SNR as SNR $= Var(\boldsymbol{O})/\sigma^2$. With $Var(x_1) = 0.25$ and $Var(\text{noise}) = 1$:

$$Var(f(x)) = Var(\beta_1 \cdot x_1) = \beta_1^2 \cdot 0.25 \Rightarrow \beta_1 = \sqrt{SNR}/0.5$$

Since we have an output $\boldsymbol{O}$ that is linear in $X_1, ..., X_5$ we can calculate ground-truth SHAP values. First, for SNR=0, we set $\boldsymbol{\beta} = \boldsymbol{0}$, which is analogous to the no-effect case in [6]. Second, we set $\boldsymbol{\beta}$ to $(1, 0, 0, 0, 0)^T$ so that only the first variable has an effect on the outcome $\boldsymbol{O}$.

For each of the generated datasets, we fitted a random forest (R package *ranger*) and gradient boosted trees (R package *xgboost*). We then calculated SHAP values with TreeSHAP (R package *treeshap*) and KernelSHAP (R package *shapr*). We ran the same data-generating process 100 times, computing SHAP values for every observation at each repetition. Further, we calculated the SHAP

variable importance by calculating the mean absolute SHAP values over the instances in the dataset [13].

Figure 1 shows the results for the no-effect case (SNR=0). As we can see, SHAP importance values exhibit a very similar behavior to uncorrected impurity importance in Strobl et al. [6], where the importance of the first continuous variable X_1 is the highest and the categorical variables are more important the more categories they have. We see the same behvaior with random forests and xgboost and both with KernelSHAP and TreeSHAP.
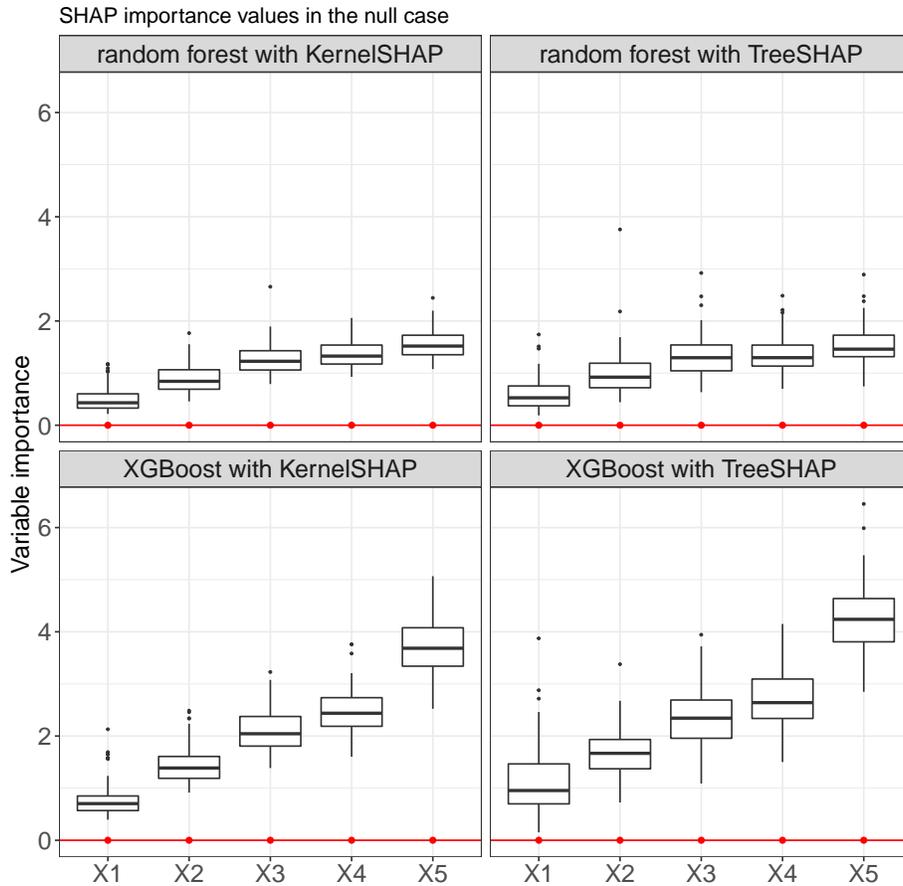
Fig. 1. Boxplots of SHAP variable importance for the no-effect case with SNR=0 with random forest and xgboost, KernelSHAP and TreeSHAP. The red dots correspond to the ground-truth SHAP values from the data-generating linear model, the red line is at 0.

Figure 2 shows the result for the effect case with $\boldsymbol{\beta} = (1, 0, 0, 0, 0)^T$ and increasing SNR. We see that for very low SNR (SNR = 0.01), X_1 receives the lowest variable importance on average, even though it is the only feature with a true effect. For a higher SNR (SNR = 0.1), the bias is still visible but X_1 can be clearly distinguished from $X_2, ..., X_5$. For a high SNR (SNR = 1), the bias is almost negligible. In the replicates of the experiment with SNR = 1 the SHAP importance value for X_1 was always the highest SHAP value out of the five variables.

Figure 10 (Appendix) shows boxplots of SHAP importance values for a neural network trained on the same data. The results confirm the fact that the bias is inherent to tree-based structures and is not necessarily exhibited by other methods.
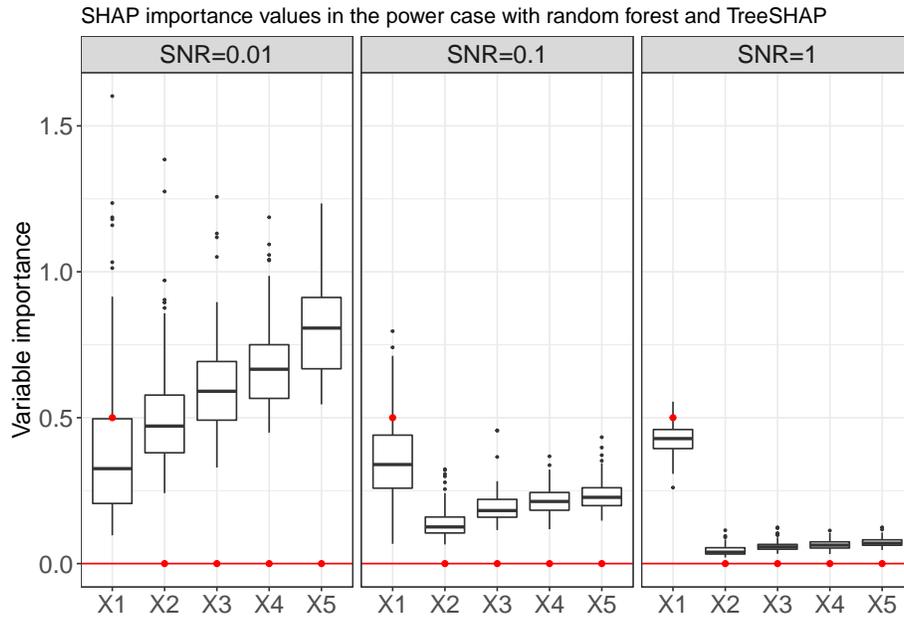

Fig. 2. Boxplots of SHAP variable importance for the effect case with $\boldsymbol{\beta} = (1, 0, 0, 0, 0)^T$ and SNR set to 0.01, 0.1 and 1, with random forest and TreeSHAP. The red dots correspond to the ground-truth SHAP values from the data-generating linear model, the red line is at 0. Similar plots for xgboost and KernelSHAP can be found in the Appendix.

3 Effect of Model Tuning

We are interested in the effect of optimally tuning Random Forests on the observed bias in SHAP importance values. In this section, we present results from a grid search aimed at finding those values for `max_depth` and `mtry`, which minimize the log loss or MSE on a hold-out set. Both for synthetic and real data sets (see Section 4) we find that the observed bias is inevitably due to a sub-optimally tuned model.

We follow the same procedure as in Section 2, but omit the often studied *null case* ($\beta_1 = 0$) since clearly the optimal model would just be a mean prediction. Instead we focus on the so-called *power case* ($\beta_1 > 0$) and tune the model parameters for various signal to noise ratios.

Figure 3 demonstrates that the out-of-bag loss is smallest for `max_depth` = 1 and `mtry` = 5, which is not surprising: if only the binary predictor x_1 is informative, the best model would be a stump with no randomization in the columns (mtry=5) which is confirmed by the graphs. In this light, the observed bias in the simulated data by Strobl et al. [6] could be viewed simply as an artefact of "throwing" an overly complex, non-tuned model at a data set with an extremely simple structure.
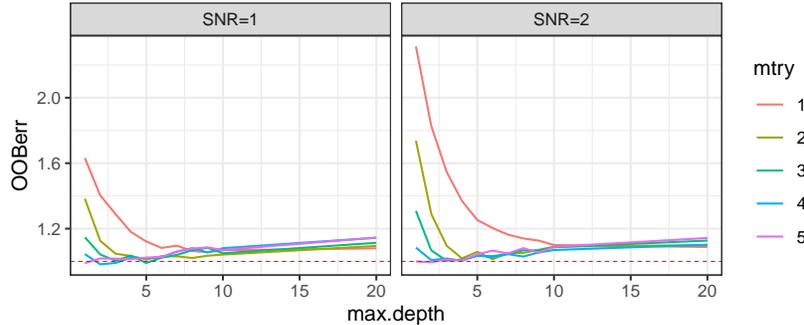
Fig. 3. Simulated data ($n = 1000$): Out-of-bag (OOB) error as a function of `max_depth` and `mtry` for SNR=1,2 respectively. The optimal RF parameters are seen to be $max\_depth = 1$ and $mtry = 5$, consistent with the obvious fact that a tree stump would be the ideal model. The Bayes error ($= 1$) is added as a brown dashed horizontal line.

A similar story emerges from the effect of model tuning on SHAP values as illustrated in Figure 4: the (global) SHAP values for the non-informative features show the familiar biased pattern - the dependence on feature cardinality $x_5 > x_4 > x_3 > x_2 > x_1$ as well as the overall increase with `max_depth`. But strikingly, for `max_depth` = 1 and `mtry` = 5, the bias completely disappears and the only non-zero SHAP scores arises for x_1. The expected/true SHAP value shown in Figure 4 can be computed as follows: The overall expected value of y
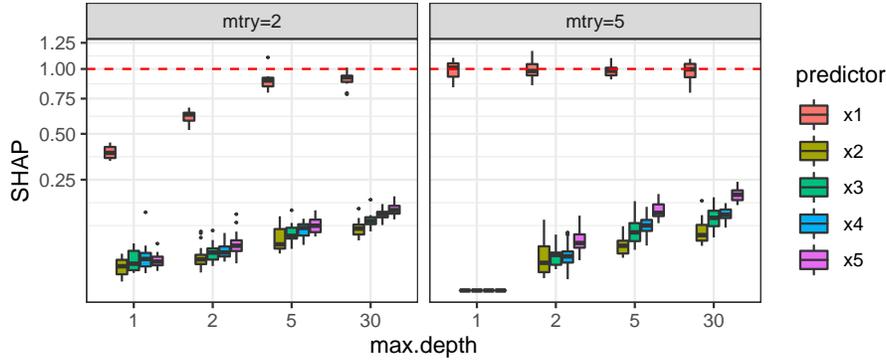
is

$$E(y) = \beta_1 E(x_1) = 1.5 \cdot \beta_1 = 1.5 \cdot \sqrt{\text{SNR}}/0.5 = 3 \cdot \sqrt{\text{SNR}}.$$

Hence

$$\phi_2 = \pm\sqrt{\text{SNR}} \Rightarrow |\phi_2| = \sqrt{\text{SNR}},$$

since $\beta_1 x_1$ takes on the values $2\sqrt{\text{SNR}}$ and $4\sqrt{\text{SNR}}$, respectively.

Fig. 4. Simulated data ($n = 200$): Global SHAP distributions (100 repeated simulations) as a function of `max_depth` for `mtry`=2,5 respectively (SNR=1). At `max_depth`=1 and `mtry`=5 (rightmost panel), the SHAP scores seem entirely unbiased: zero for non-informative features and centered around the expected/true SHAP value for x_1 which is added as a red dashed horizontal line. Note the square-root y-scale which greatly amplifies the small SHAP values for the non informative predictors.

We repeated these simulations for a sequence of decreasing SNRs and noticed marked deviation from this unbiasedness only for very low values of SNR< 0.08. In comparison, even a linear model fails to reliably identify the estimated slope β_1 as significantly different from zero for values of SNR< 0.05. So again, we conclude from this subsection that the well publicised bias seems to be of concern only in low SNR situations with poorly tuned models.

4   Real Data Analysis

The cervical cancer data set contains indicators and risk factors for predicting whether a woman will get cervical cancer. The features include demographic data (such as age), lifestyle, and medical history. The data can be downloaded from the UCI Machine Learning repository and was used in [13] to illustrate the utility of SHAP values to yield feature importance rankings, as illustrated in Figure 11 in the Appendix. As there appears to be a general consensus/expectation that random forests rarely overfit [14–16] it is frequently common practice to accept its default parameters and not spend more resources in tuning the model, which

is the case for the results in Figure 11. On closer inspection though, these feature attributions may reflect a fair degree of overfitting: Figure 5 suggests that the optimal model is simply the baseline mean prediction and that more complex ones fare worse on validation data. We speculate that this unexpected result is due to the extreme imbalance of the outcome ($\hat{y} = 0.94$).
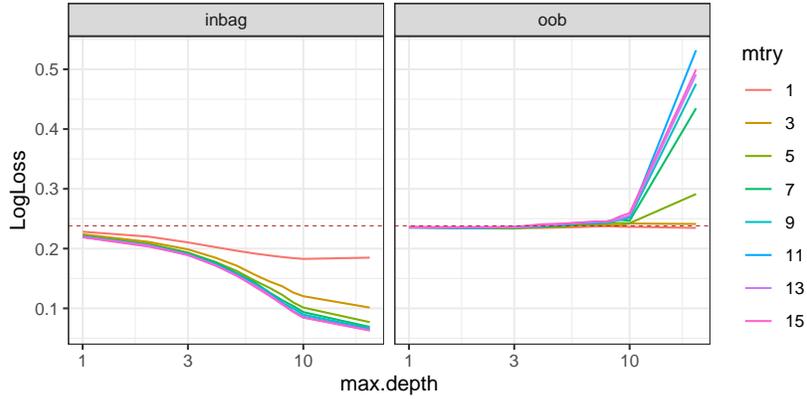
Fig. 5. Cervical cancer data: Inbag versus out-of-bag (oob) loss as a function of the two main tuning parameters of random forests. Somewhat surprisingly, the validation loss is highly insensitive to the exact choices of max-depth and mtry and is in fact minimized by the lowest complexity model, tree stumps. The baseline (mean predictions) log loss is overlaid as a dashed brown line.

To study how the bias affects SHAP values on real data, we added permuted versions of the original features, fitted random forests with default parameters and calculated TreeSHAP variable importance values. We applied this procedure to the cervical cancer (see above) and Boston housing data sets. The Boston housing dataset contains various measures related to housing for the census tracts of Boston recorded in the 1970 census. The dataset was first published in Harrison et al.[17] Figure 6 shows the results on both data sets: For the cervical cancer data, half of the top ten features are permuted ones, the two top permuted features are the ones with highest cardinality (age and hormonal contraceptives years) and for some features the permuted version even ranks higher than the original (e.g. age). Thus, the top ten feature selection appears to be mostly random, with a tendency to rank high-cardinality features high. On the other hand, on the boston housing data, all top ten features are original features. These results confirm our findings from the simulation studies: The bias towards high-cardinality features is notable in low-SNR settings (cervical cancer) but does not affect high-SNR settings (boston housing) much.
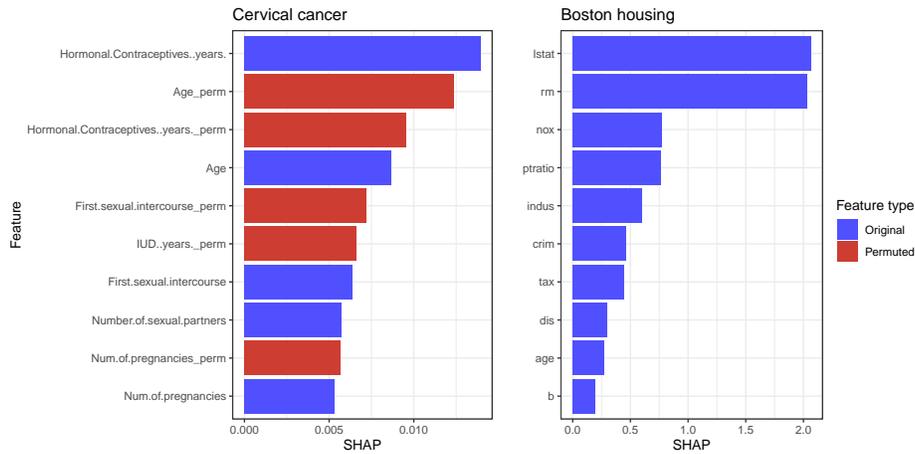
Fig. 6. SHAP variable importance (top 10) for the cervical cancer and boston housing data sets with original and permuted features. Permuted features are permuted versions of the original features in the data.

5   Related Work

Strobl et al. [6] showed that variables with more categories have a higher Gini importance. Permutation importance was also shown to be unbiased but variables with more categories have higher variance which leads to higher false positive rates for global importance measures. The authors argued that the bias is inherent to the tree structure. Many solutions have been proposed to address this bias, often by modifying the tree-building algorithm: Strobl et al. [6] suggest using conditional inference trees or forests [18] and sub-sampling without replacement to solve the problem. Wright et al. [8] make use of maximally selected rank statistics to avoid the bias, whereas Loh & Shih [19] use ANOVA F-statistics.

Loecher [20] showed that SHAP values exhibit the bias just like Gini and permutation importance in the null and power case for random forests. Adler & Painsky [12] extended the result by showing that gradient boosting machines exhibit the same bias in importance measures for Gini, permutation and SHAP importance. SHAP is originally a local measure of variable importance. To use SHAP as a global importance measure we are forced to first calculate SHAP values for each individual observation before taking the mean of the absolute value of SHAP values for each variable across observations [13]. An alternative method to calculate global SHAP values is SAGE [21], which estimates Shapley values for each variable to explain the reduction in the risk function induced by adding this variable to a coalition of variables instead of explaining each model prediction individually. Sutera et al. [22] shows that SAGE applied to tree-based models is very similar to the Gini importance. Further methods to calculate global SHAP values have been proposed [23–25]. However, as argued above the

bias is due to the underlying tree-structure and thus, all of these methods inherit it, as long as the tree-building algorithm is unchanged.

Through comprehensive experiments, Yasodhara et al. [26] evaluate both the accuracy and stability of estimated global feature importance scores such as SHAP and Gini and report disappointingly low correlations with the true feature rankings even in the absence of added noise. When inputs or models are perturbed, the correlations drop even lower. The authors did not investigate the influence of uninformative variables of varying cardinality. The stability of local explainability methods has been studied by Alvarez et al. [27] who show that LIME [28] and (Kernel) SHAP [1] lack stability for complex black-box models. Perhaps the most interesting take-away from these studies is the possible coexistence of stable predictions with fragile explanations that change drastically in response to the perturbations.

6   Discussion

SHAP variable importance measures show a bias when trying to infer an inherent relationship in the data similar to the bias observed by Strobl et al.[6]. The bias towards high-entropy variables, i.e. variables with more categories and continuous variables, is present not just for random forests, but also for trees and XGBoost trees, suggesting that this bias is due to the tree structure itself, as was put forward in [6]. We showed that SHAP variable importance inherits this bias but that the signal to noise ratio (SNR) plays an important role: In low-SNR settings, the bias is evident, up to the point where a variable with many categories and no effect on the outcome can have a higher SHAP importance value than a variable with few categories but an effect on the outcome. On the other hand, in high-SNR settings, the bias is still there but its magnitude is vanishingly small, compared to true effects and thus of no major concern for the practitioner.

The real world examples show the importance of SNR to determining the importance of the bias. On the cervical cancer dataset with low SNR, some of the variables with the highest SHAP values were permuted copies with high cardinality. On the Boston housing dataset with high SNR, the permuted features were not in the top ten most important features by SHAP importance. We also showed that in the power case, careful tuning of the hyperparameters of a random forest can effectively eliminate the bias. Picking the `mtry` and maximum tree depth combination with the lowest validation error, we end up with `mtry=5` and `max_depth=1`. If we let the model have a maximum depth of 1 and let it choose between all five variables at each split, the bias disappears.

It is important to note that the bias studied in this paper can only be considered a bias when we use variable importance measures to infer which variable is important for some underlying relationship in the data [29]. If we are instead interested in understanding the tree-based model itself, than this "bias" in fact corresponds to a true effect, as the tree-based structure does in fact use variables with more categories more often and closer to the top of trees than

variables with fewer categories. This leads to an important distinction between biases: A variable importance measure can be biased for identifying which variable is important in determining the true outcome and it can also be biased for identifying which variable is important in determining the predicted outcome, i.e. the model output. If the model to be explained is not a perfect rendition of the underlying relationship in the data, then any variable importance measure will exhibit one of these biases. Thus, when using a variable importance measure one should always consider what the measure should explain, the model output or the real-world outcome.

In conclusion, recommend practitioners to consider the bias towards high-entropy features when interpreting SHAP-based variable importance. If low prediction performance on validation data indicates a low signal to noise ratio, the bias can affect feature rankings and careful hyperparameter tuning is particularly important. A strategy to detect the impact of the bias in sensitivity analyses could be to add permuted feature copies as *fake* feature to the data. One step further, one could also apply a correction method, as e.g. proposed by Nembrini et al. [9].

References

[1]   Lundberg, S. M. and Lee, S.-I. "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems.* Vol. 30. 2017.

[2]   Shapley, L. S. "A value for n-person games". In: *Contributions to the Theory of Games (AM-28), Volume II.* Ed. by H. W. Kuhn and A. W. Tucker. Princeton University Press, 2016, pp. 307–318.

[3]   Štrumbelj, E. and Kononenko, I. "Explaining prediction models and individual predictions with feature contributions". *Knowledge and Information Systems* 41.3 (2014), pp. 647–665.

[4]   Lundberg, S. M., Erion, G. G., and Lee, S.-I. "Consistent individualized feature attribution for tree ensembles". *arXiv preprint arXiv:1802.03888* (2018).

[5]   Breiman, L. et al. *Classification and Regression Trees.* CRC press, 1984.

[6]   Strobl, C. et al. "Bias in random forest variable importance measures: Illustrations, sources and a solution". *BMC Bioinformatics* 8.1 (2007), pp. 1–21.

[7]   Boulesteix, A.-L. et al. "Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations". *Briefings in Bioinformatics* 13.3 (2012), pp. 292–304.

[8]   Wright, M. N., Dankowski, T., and Ziegler, A. "Unbiased split variable selection for random survival forests using maximally selected rank statistics". *Statistics in Medicine* 36.8 (2017), pp. 1272–1284.

[9]   Nembrini, S., König, I. R., and Wright, M. N. "The revival of the Gini importance?" *Bioinformatics* 34.21 (2018), pp. 3711–3718.

[10] Loecher, M. "From unbiased MDI feature importance to explainable AI for trees". *arXiv preprint arXiv:2003.12043* (2020).

[11] Loecher, M. "Debiasing MDI feature importance and SHAP values in tree ensembles". In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer. 2022, pp. 114–129.

[12] Adler, A. I. and Painsky, A. "Feature importance in gradient boosting trees with cross-validation feature selection". *Entropy* 24.5 (2022), p. 687.

[13] Molnar, C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2019.

[14] Breiman, L. "Random forests". *Machine Learning* 45.1 (2001), pp. 5–32.

[15] Probst, P., Wright, M. N., and Boulesteix, A.-L. "Hyperparameters and tuning strategies for random forest". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.3 (2019), e1301.

[16] Probst, P., Boulesteix, A.-L., and Bischl, B. "Tunability: Importance of hyperparameters of machine learning algorithms". *Journal of Machine Learning Research* 20.1 (2019), pp. 1934–1965.

[17] Harrison, D. and Rubinfeld., D. "Hedonic housing prices and the demand for clean air". *Journal of Environmental Economics and Management* 5 (1978), pp. 81–102.

[18] Hothorn, T., Hornik, K., and Zeileis, A. "Unbiased recursive partitioning: A conditional inference framework". *Journal of Computational and Graphical Statistics* 15.3 (2006), pp. 651–674.

[19] Loh, W.-Y. and Shih, Y.-s. "Split selection methods for classification trees". *Stat Sinica* 7 (July 1999).

[20] Loecher, M. "Unbiased variable importance for random forests". *Communications in Statistics - Theory and Methods* 51.5 (2020), pp. 1–13.

[21] Covert, I., Lundberg, S. M., and Lee, S.-I. "Understanding global feature contributions with additive importance measures". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 17212–17223.

[22] Sutera, A. et al. "From global to local MDI variable importances for random forests and when they are Shapley values". *Advances in Neural Information Processing Systems* 34 (2021), pp. 3533–3543.

[23] Casalicchio, G., Molnar, C., and Bischl, B. "Visualizing the feature importance for black box models". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2018, pp. 655–670.

[24] Frye, C., Rowat, C., and Feige, I. "Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability". *Advances in Neural Information Processing Systems* 33 (2020), pp. 1229–1239.

[25] Williamson, B. and Feng, J. "Efficient nonparametric statistical inference on population feature importance using Shapley values". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 10282–10291.

[26] Yasodhara, A. et al. "On the trustworthiness of tree ensemble explainability methods". In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer. 2021, pp. 293–308.

[27]   Alvarez-Melis, D. and Jaakkola, T. S. "On the robustness of interpretability methods". *arXiv preprint arXiv:1806.08049* (2018).

[28]   Ribeiro, M. T., Singh, S., and Guestrin, C. ""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.* 2016, pp. 1135–1144.

[29]   Chen, H. et al. "True to the model or true to the data?" *arXiv preprint arXiv:2006.16234* (2020).
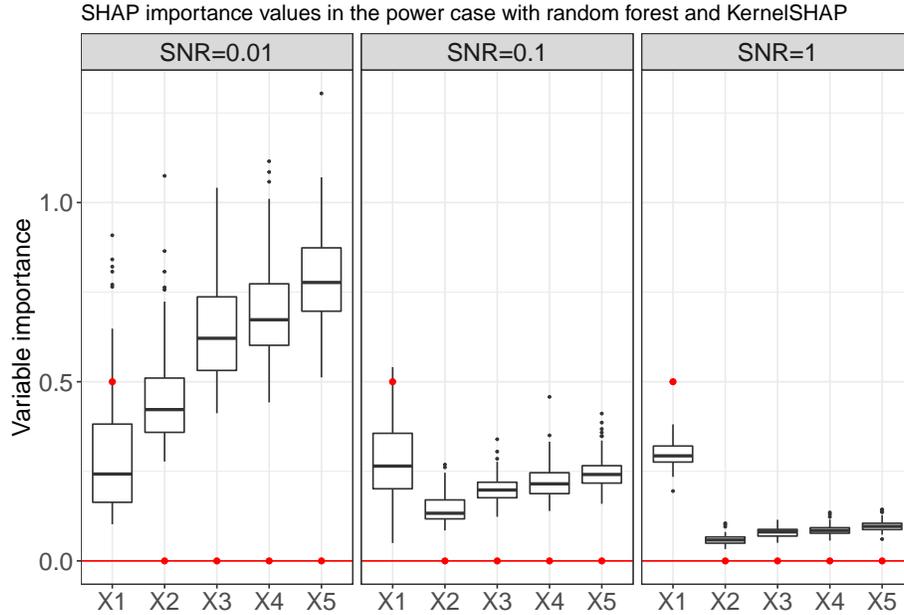
Fig. 7. Boxplots of SHAP variable importance for the effect case with $\boldsymbol{\beta} = (1, 0, 0, 0, 0)^{T}$ and SNR set to 0.01, 0.1 and 1, with random forest and KernelSHAP. The red dots correspond to the ground-truth SHAP values from the data-generating linear model, the red line is at 0.

A   Appendix

SHAP importance values in the power case with XGBoost and TreeSHAP
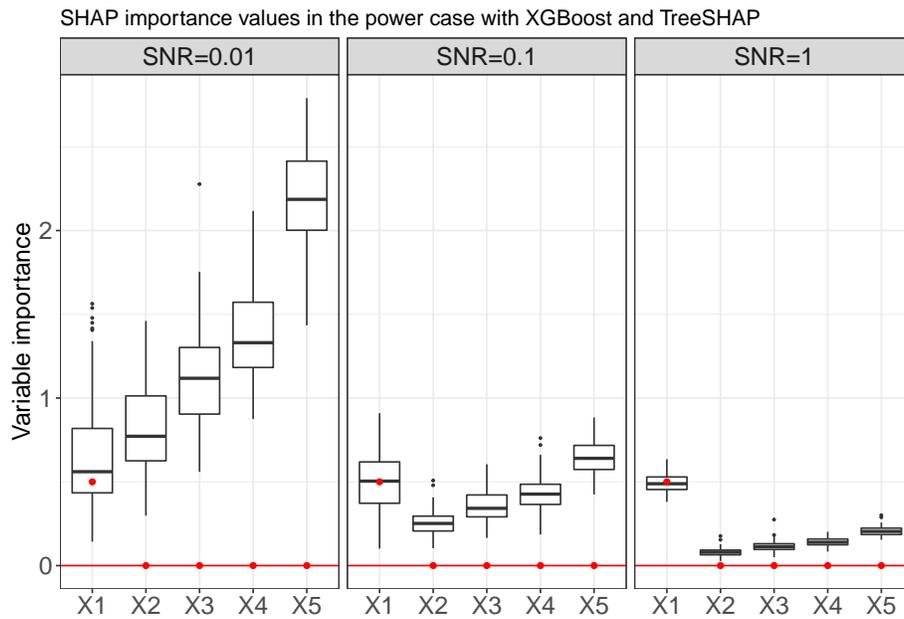
Fig. 8. Boxplots of SHAP variable importance for the effect case with $\boldsymbol{\beta} = (1, 0, 0, 0, 0)^T$ and SNR set to 0.01, 0.1 and 1, with XGBoost and TreeSHAP. The red dots correspond to the ground-truth SHAP values from the data-generating linear model, the red line is at 0.

SHAP importance values in the power case with XGBoost and KernelSHAP

Fig. 9. Boxplots of SHAP variable importance for the effect case with $\boldsymbol{\beta} = (1, 0, 0, 0, 0)^T$ and SNR set to 0.01, 0.1 and 1, with XGBoost and KernelSHAP. The red dots correspond to the ground-truth SHAP values from the data-generating linear model, the red line is at 0.

SHAP importance values for each variable in a two−layer neural network
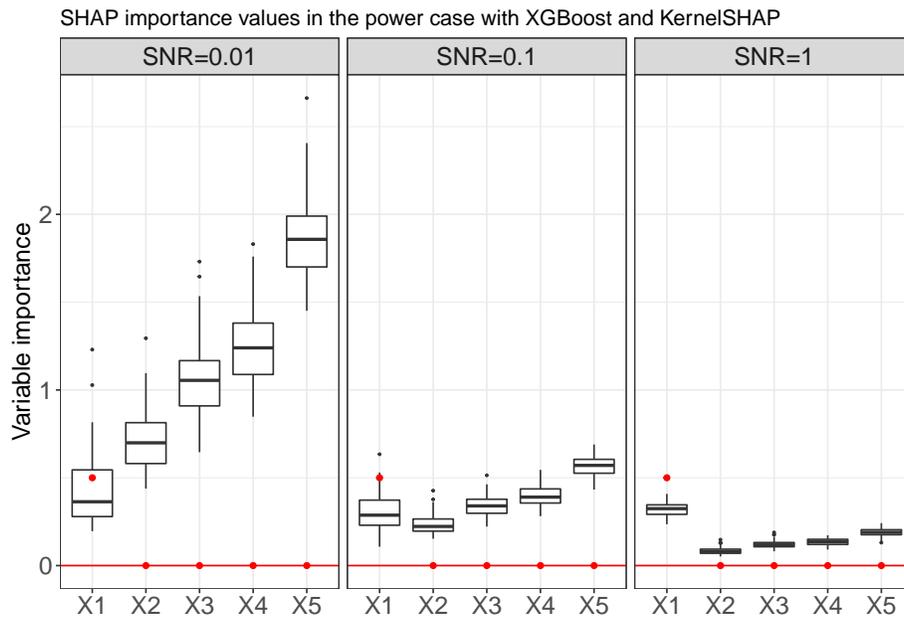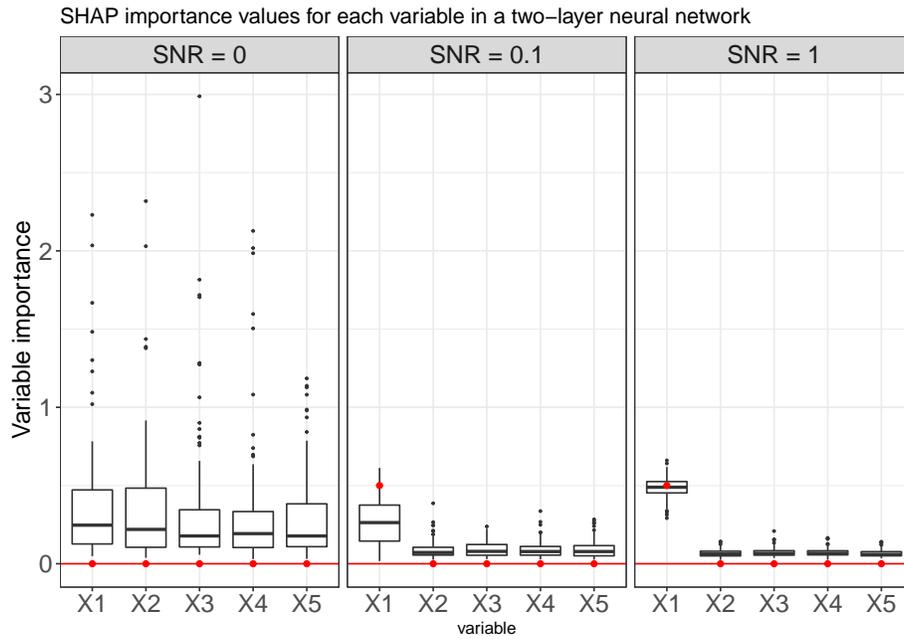
Fig. 10. Boxplots of SHAP variable importance for a neural network with two layers of 64 units with relu activation, Adam iptimization and a linear output layer. The first plot is for the no effect case and the second and third plots are for the effect case with $\boldsymbol{\beta} = (1, 0, 0, 0, 0)^{T}$ and SNR set to 0.1 and 1. The red dots correspond to the ground-truth SHAP values from the data-generating linear model, the red line is at 0
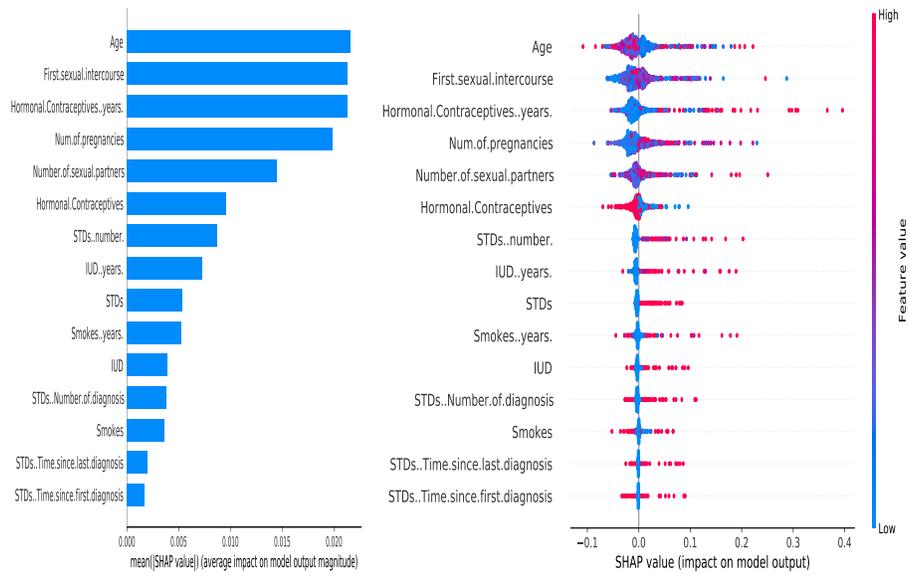
Fig. 11. left: SHAP feature importance measured as the mean of the absolute SHAP values. **right**: SHAP "beeswarm" plot showing the distributions of SHAP values per feature color coded by the sign and magnitude of the corresponding feature value. Low number of years on hormonal contraceptives reduce the predicted cancer risk, a large number of years increases the risk. (Random Forest with default parameter settings, ntree=100)