


CORRESPONDENCE

Addressing bias in feature importances derived from XGBoost. Comment on *Br J Anaesth* 2024;133:351–9

Yoshiyasu Takefuji 

Faculty of Data Science, Musashino University, Tokyo, Japan

E-mail: takefuji@keio.jp

Keywords: bias; feature importance; machine learning; SHAP analysis; statistical methods; XGBoost

Editor—In their recent study of opioid use in children, Atias and colleagues¹ utilised SHAP (SHapley Additive exPlanations) analysis in conjunction with eXtreme Gradient Boosting (XGBoost) to identify influential variables such as the number of diagnoses, medical imaging, laboratory tests, and the type of opioid used. However, the application of SHAP analysis is not ideal and is limited because of inherent biases in the feature importances derived from XGBoost, which can lead to misleading conclusions. This model-specific nature implies that different models yield distinct sets of feature importances, even when the true associations between the target and features are calculable. To avoid biased feature importances, it is recommended to rely on statistical methods that assess genuine associations, such as Spearman's correlation with P-values.

The recommendation to use Spearman's correlation with P-values is primarily because of its nonparametric nature, which makes it suitable for analysing data that might not meet the assumptions of a normal distribution. This method is particularly valuable when dealing with ordinal data or when the relationship between variables is not linear. It is important to note that these methods impose specific assumptions regarding the underlying data distributions. For instance, Student's t-test requires normally distributed data, whereas the Wilcoxon rank sum test is used for comparing two independent samples under certain conditions. Fisher's exact test, although useful for small sample sizes, applies specifically to categorical data organised in a contingency table. By highlighting Spearman's correlation, this paper aims to emphasise a robust approach that can be more broadly applicable across varied datasets without stringent assumptions. This flexibility allows researchers to capture genuine associations in their analyses, making it strongly recommended for preliminary investigations.

Biases in feature importance metrics can arise from misinterpretations related to SHAP values, specifically in the

context of the function $\text{explain}=\text{SHAP}(\text{model})$. An extensive body of literature, comprising over 100 articles, has investigated biases in feature importance derived from machine learning models, particularly XGBoost. It is crucial to understand that SHAP values are inherently connected to the underlying machine learning model, in this instance, XGBoost. As a result, SHAP values can inherit and even exacerbate the biases present in the model, underscoring the need for careful interpretation when evaluating feature importance.

Although many researchers, including Atias and colleagues, utilise SHAP analysis, machine learning models such as XGBoost can produce biased feature importances,^{2–6} so unbiased, machine learning-independent methods that provide true associations are preferred and should be used. The primary goal of machine learning is to predict the target accurately; however, feature importances derived from these models are merely byproducts. Each model employs different algorithms for calculating feature importances, a phenomenon known as model-specific bias.^{2–6}

Different machine learning models employ distinct methodologies for calculating feature importances, leading to varying degrees of bias. For instance, decision tree-based algorithms, such as XGBoost, build ensembles of trees sequentially, with each tree attempting to rectify the errors of its predecessors. Although XGBoost is highly effective for predictive tasks, the inherent characteristics of this approach can introduce biases in the reported feature importance values. In addition to XGBoost, other models, such as linear regression and certain ensemble methods, can also exhibit biases depending on their underlying assumptions and structures.

Although several bias mitigation techniques exist, none can entirely eliminate these biases from feature importances. Given this context, it is important to use Spearman's correlation with P-values to assess true associations between the target variable and features. This method accounts for

nonparametric relationships and offers a robust means of evaluating feature relevance, thereby reducing the reliance on potentially biased feature importance values.

Machine learning models such as XGBoost can generate feature importance biases for several reasons. Firstly, the concept of model specificity plays a significant role. Different algorithms have their unique methodologies for calculating feature importance, which can result in considerable variation across models. For instance, XGBoost employs a tree-based approach that evaluates each feature's contribution based on criteria such as gain, cover, or the frequency of splits. This reliance on specific metrics can lead to biased estimations of importance, which might not accurately reflect the actual relationships within the data. Secondly, interaction effects can complicate the interpretation of feature importance. Tree-based models, including XGBoost, tend to capture interactions among features, meaning a feature's importance can be artificially inflated if it interacts with another feature. In contrast, a feature can appear less important when assessed independently. Such interactions can obscure the true contributions of individual features, making it challenging to gauge their real impact.

Another contributing factor is the presence of correlated features. When features are highly correlated, the model might assign importance to one feature over another on an arbitrary basis. This phenomenon can lead to misleading conclusions regarding which features genuinely drive predictions, as multiple features might provide redundant information. When features are correlated or collinear, the model's assignment of importance to individual features can appear somewhat arbitrary because of the interconnected nature of these variables. In essence, when two or more features convey similar information regarding the target variable, the model might disproportionately attribute importance to one feature over the others based on the order in which the features are added during the training process, especially in algorithms such as decision trees and ensemble methods (e.g. XGBoost).

For example, in decision tree algorithms, the importance of a feature is often derived from the reduction in impurity (e.g. Gini impurity or entropy) that the feature provides when it is used to split the data. If two correlated features are available, the model might select one feature for the initial splits, especially if it happens to yield a slightly better impurity reduction at that point. As the tree grows, the first feature chosen can dominate the importance ranking, overshadowing the contributions of the other correlated feature(s). This can lead to a misleading interpretation where one feature is deemed significantly more important than another, even though both features are fundamentally providing similar insights into the underlying patterns in the data. This phenomenon can result in instability in the feature importance rankings when the model is retrained with slight variations in the data or other hyperparameters. Additionally, the assignment of importance can differ significantly across different models, or even different configurations of the same model, depending on how they handle correlated features.

Overfitting is also a critical concern. XGBoost, like other sophisticated models, can overfit the training data, resulting in the assignment of high importance to features that seem relevant solely within the context of the training set. Such features might not hold true across different datasets, creating a biased perception of importance that lacks generalisability.

Additionally, the lack of statistical validation in assessing feature importance is noteworthy. Traditional statistical

methods, such as P-values and confidence intervals, offer frameworks for evaluating the reliability of estimates. In contrast, feature importance scores derived from machine learning models often lack such statistical backing, leading to uncertainties regarding their validity. Therefore, what might seem like a significant feature might not be statistically supported.

The bias–variance trade-off further complicates matters. As models become more complex to reduce bias by capturing intricate patterns in the data, they can inadvertently increase variance. This increase can cause the importance of some features to fluctuate significantly across different samples or cross-validation folds, resulting in inconsistent feature importance assessments.

Lastly, feature engineering can also introduce biases into feature importance. The way features are constructed or transformed can skew their importance scores. For example, aggregating features or applying nonlinear transformations may distort the relationships the model assigns to them. Recognising these biases is crucial for researchers, as it prompts a cautious interpretation of feature importance derived from XGBoost and similar machine learning models. To address potential biases, employing complementary statistical methods to unveil true associations can provide a more accurate understanding of feature contributions.

It is crucial to understand that feature engineering can introduce biases into feature importance calculations if not performed thoughtfully. Although comparing SHAP values within a model can provide insights into relative feature importance, these comparisons can still carry biases derived from both the underlying model and the feature engineering process itself. For feature engineering to be effective and yield valid insights, it is essential that the base model is unbiased. When biases are present in the model, any engineered features might inadvertently amplify these biases rather than mitigate them. This is particularly concerning when using SHAP values, as they are intrinsically dependent on the model being analysed. Because SHAP values are computed based on how the model weights the contributions of individual features, any biases in the model will be reflected in the SHAP calculations, potentially skewing the interpretation of feature importance.

Therefore, although relative comparisons of SHAP values among features can shed light on their importance within the context of a specific model, it remains critically important to be cautious of the potential biases that can affect these values. Approaches aimed at minimising bias are essential for accurately capturing true associations between features and the target variable. No method, SHAP included, can completely eliminate the effect of biases inherent in the model or introduced through the feature engineering process. By ensuring rigorous validation and critical evaluation of both the feature engineering approach and model performance, researchers can work towards more reliable interpretations of feature importance. This is not meant to discourage machine learning for prediction, but rather to encourage analysis of the true associations between the target and features without biases using statistical methods such as Spearman's correlation with P-values.^{7–9}

In conclusion, although Atias and colleagues¹ leveraged SHAP analysis alongside XGBoost to identify key variables in healthcare, the inherent biases in feature importance derived from this model raise significant concerns about the validity of their findings. The model-specific nature of feature

importance can lead to misleading conclusions, as different algorithms can yield disparate importance scores despite potentially calculable true associations. To enhance the reliability of insights, it is essential to employ unbiased statistical methods, such as χ^2 -tests and Spearman's correlation, which provide a more accurate assessment of variable relevance. Ultimately, recognising and addressing these biases will lead to a clearer understanding of the relationships between features and the target in machine learning applications.

Declaration of interest

The author declares that they have no conflict of interest.

References

1. Atias D, Tuttnauer A, Shomron N, Obolski U. Prediction of sustained opioid use in children and adolescents using machine learning. *Br J Anaesth* 2024; **133**: 351–9
2. Thakur D, Biswas S. Permutation importance-based modified guided regularized random forest in human activity recognition with smartphone. *Eng Appl Artif Intell* 2024; **129**, 107681
3. Barton-Henry K, Wenz L, Levermann A. Decay radius of climate decision for solar panels in the city of Fresno, USA. *Sci Rep* 2021; **11**: 8571
4. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy* 2021; **23**: 18
5. Esser-Skala W, Fortelny N. Reliable interpretability of biology-inspired deep neural networks. *npj Syst Biol Appl* 2023; **9**: 50
6. Chen J, Ooi LQR, Tan TWK, et al. Relationship between prediction accuracy and feature importance reliability: an empirical and theoretical study. *Neuroimage* 2023; **274**, 120115
7. Yu H, Hutson AD. A robust Spearman correlation coefficient permutation test. *Commun Stat Theory Methods* 2024; **53**: 2141–53
8. Eden SK, Li C, Shepherd BE. Nonparametric estimation of Spearman's rank correlation with bivariate survival data. *Biometrics* 2022; **78**: 421–34
9. Jiang I, Zhang X, Yuan Z. Feature selection for classification with Spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets. *Expert Syst Appl* 2024; **249**(B), 123633

doi: 10.1016/j.bja.2024.11.033