Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Beyond XGBoost and SHAP: Unveiling true feature importance

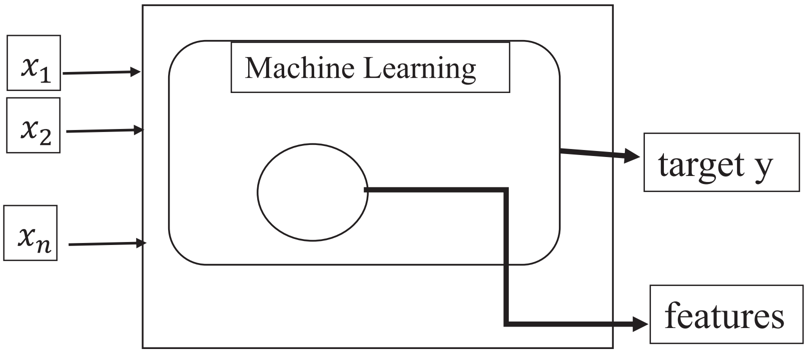
Yoshiyasu Takefuji [1]

Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan

HIGHLIGHTS

- XGBoost may induce biased feature importances due to model specific nature.
- SHAP values may inflate feature importance scores due to model biases.
- Absence of ground truth complicates feature importance validation efforts.
- Robust statistical methods help improve reliability of machine learning analyses.
- Understanding biases is crucial for accurate interpretations in research.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Machine learning
XGBoost
SHAP values
Feature importance
Statistical methods for validation

ABSTRACT

This paper outlines key machine learning principles, focusing on the use of XGBoost and SHAP values to assist researchers in avoiding analytical pitfalls. XGBoost builds models by incrementally adding decision trees, each addressing the errors of the previous one, which can result in inflated feature importance scores due to the method's emphasis on misclassified examples. While SHAP values provide a theoretically robust way to interpret predictions, their dependence on model structure and feature interactions can introduce biases. The lack of ground truth values complicates model evaluation, as biased feature importance can obscure real relationships with target variables. Ground truth values, representing the actual labels used in model training and validation, are crucial for improving predictive accuracy, serving as benchmarks for comparing model outcomes to true results. However, they do not ensure real associations between features and targets. Instead, they help gauge the model's effectiveness in achieving high accuracy. This paper underscores the necessity for researchers to recognize biases in feature importance and model evaluation, advocating for the use of rigorous statistical methods to enhance the reliability of analyses in machine learning research.

This paper outlines the fundamental principles of machine learning and illustrates the application of XGBoost in conjunction with SHAP values, providing researchers with essential tools to navigate common pitfalls in their analyses.

Many researchers, lacking a solid understanding of the fundamental principles of machine learning, often apply methodologies without fully grasping their implications. The primary goal of machine learning is to accurately predict target outcomes, and this is typically achieved

E-mail address: takefuji@keio.jp.

[1] **ORCID:** 0000–0002-1826–742X

https://doi.org/10.1016/j.jhazmat.2025.137382
Received 5 November 2024; Received in revised form 16 January 2025; Accepted 24 January 2025
Available online 27 January 2025

through supervised learning, which relies on ground truth values for validation. Within the realm of supervised machine learning, there are two main types: classification and regression.

Classification tasks involve predicting categorical outcomes, which are generally represented as discrete values, such as integers. Evaluation of classification accuracy includes metrics such as accuracy, sensitivity, specificity, and other validation metrics. On the other hand, regression focuses on predicting continuous real numbers. Instead of accuracy metrics, regression analysis utilizes performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared, Adjusted R-squared, and Mean Absolute Percentage Error (MAPE). These metrics quantify a model's goodness of fit, indicating how well its predictions align with the observed data. It's essential to recognize the distinction between these regression metrics and accuracy metrics used in classification tasks; while good fit often correlates with better accuracy, they serve different purposes.

In scenarios where ground truth values are unavailable, careful assessment of machine learning models becomes paramount to ensure accurate analysis. Here, the secondary goal of machine learning emerges: to calculate feature importances or uncover true associations between the target variable and the features. In unsupervised machine learning and clustering, there are no ground truth values available for validating accuracy, unlike in supervised machine learning, which relies on ground truth labels for performance assessment. The lack of ground truth values means that different models must employ distinct methodologies to calculate feature importance. As a result, feature importance varies between models, emphasizing their inherent model-specific nature. This nuance suggests that feature importances may be biased due to these model-specific characteristics.

To mitigate such biases, various methods can be employed, although none can entirely eliminate all biases in feature importance evaluations. To accurately ascertain genuine relationships between the target and features, three critical considerations must be made: the data distribution, the statistical relationships between variables, and the validity of those relationships as indicated by p-values. Given these complexities, this paper advocates for the adoption of bias-free robust statistical methods, such as Spearman's correlation and Kendall's tau, both of which are nonparametric and nonlinear approaches that incorporate p-values.

When discussing classification and regression, a key distinction lies in the type of output generated. While classification deals with categorical outcomes, regression is focused on continuous values. As such, converting real numbers into integers while preserving their significant digits is important. Despite the absence of traditional accuracy analysis in regression, evaluating the predictive accuracy of these models remains crucial.

To this end, our paper introduces a novel approach that involves transforming regression tasks into classification problems, thereby enabling the application of accuracy analysis. By ranking real numbers or converting continuous values into categorical segments, researchers can more effectively examine model performance through familiar classification metrics. Regression deals with continuous target values, typically real numbers, while classification focuses on discrete outcomes represented by integers. Despite using similar algorithms, the key distinction lies in the interpretation of results, with careful attention needed to significant digits in both cases to ensure accuracy.

Understanding the distinctions between classification and regression, along with the implications of feature importance and model-specific biases, is essential for leveraging machine learning effectively. By employing robust statistical methods and recognizing the potential pitfalls of feature importance calculations, researchers can enhance the reliability of their analyses and draw more informed conclusions.

While cross-validation, along with data splitting and shuffling, is effective for validating the accuracy of model predictions, it is not suitable for assessing the accuracy of feature importance. This distinction is crucial, as the importance of a feature does not necessarily correlate with its predictive power in isolation. Cross-validation is a powerful technique for evaluating prediction accuracy; however, in time-series analysis that relies on historical data, special precautions must be taken when applying cross-validation.

In machine learning, the choice between linear or nonlinear models, as well as parametric or nonparametric analyses, is vital for ensuring accurate results. When dealing with nonlinear and nonparametric data, applying linear models or parametric analyses can lead to significant distortions in outcomes. For instance, Principal Component Analysis (PCA) is a widely used technique that operates under linear and parametric assumptions. While PCA can effectively reduce dimensionality and identify key components of the data, it often falls short in capturing essential features in complex, real-world problems [1]. When PCA is used in the scikit-learn library, it is characterized as a linear and parametric method. In contrast, when KernelPCA is employed, it is considered a nonlinear and nonparametric technique.

This limitation occurs because PCA assumes linear relationships and may overlook non-linear interactions that are critical in identifying meaningful patterns within the data. As a result, relying solely on PCA can lead to an incomplete understanding of the underlying data structure and potentially misinform subsequent analyses. While techniques like PCA have their place in data analysis, it is essential to choose appropriate modeling approaches that align with the nature of the data. This careful consideration will lead to more meaningful interpretations and better outcomes in machine learning applications.

Qin et al. developed models aimed at predicting the rate constants for bromine atoms and dibromine radicals [2]. To identify the optimal molecular fingerprints (MFs) for constructing quantitative structure-activity relationship (QSAR) models of the rate constants for reactive bromine species (RBS), they employed SHAP analysis to evaluate how variations in the radius and length of the MFs affected model performance using the XGBoost algorithm [1]. However, their feature selection process with XGBoost was less than ideal, primarily due to significant inherent biases associated with the model's specific nature.

It is essential for researchers, including Qin et al., to differentiate the predictive capabilities of machine learning models from the relationships between target variables and their features. Understanding this distinction enhances the interpretation of results and the applicability of feature importance analyses. While machine learning primarily focuses on accurately predicting target variables, feature importance metrics are intended to elucidate the associations between these targets and the input features. Nevertheless, models like XGBoost consistently introduce biases in feature importance, which can lead to misleading conclusions [3–6]. Researchers must acknowledge that, although machine learning is a powerful tool for prediction, the feature importances derived from these models do not necessarily reflect true associations, as they are shaped by the model's inherent biases. Data reliability and the number of instances are essential for ensuring accurate training in machine learning models. The lack of ground truth values for evaluating feature importances leads to varying assessments across different models, resulting in inherently biased metrics. Over 100 peer-reviewed articles have documented the significant biases associated with feature importances derived from these models. Many researchers face challenges in accurately assessing regression performance; converting regression problems into classification tasks can improve prediction accuracy.

In regression analysis, key assessment metrics include Mean Absolute Error (MAE), which measures average absolute differences; Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), which focus on squared differences and their square root. R-squared indicates the variance explained by predictors, while Adjusted R-squared accounts for the number of predictors. Mean Absolute Percentage Error (MAPE) evaluates average percentage differences, and the Explained Variance Score measures the predictable variance in the dependent variable. While these metrics offer valuable insights into model fitting, they do not fully evaluate prediction accuracy. This paper acknowledges the utility of regression but contends that metrics such as sensitivity, specificity, and accuracy are crucial for comprehensive assessment.

Moreover, since SHAP analysis relies exclusively on machine learning models, it also inherits these biases, contributing to inaccuracies in interpretation [7]. Qin et al. also noted several inherent limitations associated with the SHAP method [2].

This paper elucidates the reasons behind the biased conclusions drawn from traditional feature selection methods and connects these biases to the algorithmic limitations that influence feature importance assessments. The lack of ground truth values in feature importance assessments from machine learning models leads to inherent biases. While cross-validation is effective in enhancing target prediction accuracy, it does not effectively address the accuracy of feature importance metrics. By analyzing common statistical tests, such as Chi-squared tests with p-values and Spearman's correlation with p-values, we reveal how reliance on these methods can lead to misinterpretations of feature significance, ultimately compromising model integrity and performance [8–11].

In the absence of ground truth values, three key elements are essential for establishing true associations between the target and features: data distribution, the statistical relationship between variables, and the validation of statistical significance through p-values. While the Chi-squared test indicates the strength of associations, Spearman's correlation provides both strength and directional information, with their respective p-values validating the significance of these relationships. It's important to note that while feature importance in machine learning models typically ranges from 0 to 1, Spearman coefficients range from −1–1, reflecting different interpretative scales.

The absence of ground truth values for validating feature importance introduces two significant challenges for XGBoost: its tree-based methodology and the presence of correlated features. XGBoost constructs an ensemble of decision trees, calculating feature importance based on the frequency of a feature's use in splits and the gain associated with each feature. Consequently, features that are frequently utilized or contribute more to information gain can dominate the importance ranking, which can be misleading, as other features may play a significant role in the model's predictions but are used less frequently. Additionally, when features are correlated, XGBoost tends to favor one feature over others in the split decision-making process. This bias can result in the misattribution of importance; a single feature might absorb the significance that should be shared among several correlated features, leading to skewed feature importance scores. As a result, understanding the true contributions of each feature can become challenging without proper validation methods in place. Because ground truth values are not available, various models—such as MLR, XGBoost, LightGBM, and CatBoost—employ different methodologies for calculating feature importance. Qin et al. demonstrated that these models yield divergent feature importance rankings, potentially leading to misleading conclusions.

XGBoost, an implementation of gradient boosted decision trees, is a highly effective machine learning algorithm known for its predictive performance. However, its design and functioning can introduce biases in feature importance metrics, particularly when combined with SHAP (SHapley Additive exPlanations) values. Understanding the algorithmic perspectives that contribute to these biases is crucial for researchers to make informed decisions about the validity of feature importance interpretations.

First, XGBoost constructs its models by sequentially adding decision trees, each trained to correct the errors of its predecessor. This boosting process inherently gives more weight to misclassified examples, which can amplify the importance of certain features based on how frequently they are used to make decisions. Consequently, features that contribute to early decisions in the tree-building process may appear disproportionately important, even if they do not substantively influence the overall predictive performance. This phenomenon occurs because the model is optimized to minimize a loss function based on the predictions it makes, and features that lead to better immediate gains can overshadow those that have significant, yet indirect, effects. As a result, SHAP values may yield inflated importance scores that do not reflect the true effects of the features when considered independently.

SHAP values offer a theoretically sound method for interpreting model predictions by quantifying the contribution of each feature to the final prediction. However, these values are calculated based on the model's structure and the interactions among features in XGBoost. Since SHAP relies on the mean over various permutations of feature values to determine each feature's contribution, it inherits the decision-making biases present in the XGBoost model itself. For instance, if a feature is primarily influential in certain regions of the feature space but not across the entire dataset, SHAP may overstate its importance if those regions are frequently encountered during model training.

Moreover, the interaction effects in decision trees can further complicate feature importance assessments. XGBoost captures complex interactions between features through its tree-structured representation. While this ability is advantageous for capturing nonlinear relationships, it can lead to misleading interpretations of feature importance when independent features interact in unexpected ways. When SHAP values are computed, they attempt to account for these interactions; however, the underlying tree's structure may still skew the perceived contributions of individual features, especially in cases where certain features dominate the interaction landscape.

A high degree of fit indicated by RMSE does not guarantee the reliability of feature importance rankings generated by XGBoost [2]. The tree-based methodology and the presence of correlated features in XGBoost mean that biases in feature importances cannot be entirely mitigated through cross-validation. While SHAP values offer valuable interpretations, they exclusively rely on the underlying model, such as XGBoost, which can lead to the inheritance and amplification of any biases found in its feature importances. Unlike target accuracy in machine learning, which can be validated against known ground truth values, feature importances lack corresponding benchmarks for validation. This absence of a reference complicates the assessment of their accuracy and reliability.

To accurately calculate true associations or genuine relationships between the target variable and features in the absence of ground truth values, three critical elements must be considered: the distribution of the data, the statistical relationships among the variables, and the validity of those relationships as indicated by p-values. It is essential to first conduct Variance Inflation Factor (VIF) analysis to assess and mitigate collinearity among features; high VIF values can inflate the significance of certain predictors and obscure their true relationships with the target variable [12,13]. By addressing collinearity and ensuring a robust statistical framework, we can enhance the reliability of our feature importance assessments and draw more meaningful conclusions about the underlying data dynamics.

In conclusion, the distinctions between classification and regression, alongside the inherent biases in feature importance calculations in machine learning, are pivotal for researchers to comprehend. The process by which XGBoost builds models—concentrating on correcting previous errors—can skew feature importance towards those that contribute early in decisions. Moreover, the complexities of SHAP values, influenced by the interactions among features and the model's structure, complicate accurate assessments of feature contributions. While transforming regression tasks into classification tasks enhances performance evaluation, it is crucial to recognize that a high degree of fit indicated by RMSE does not guarantee reliability in feature importance rankings. Without ground truth for validation, the assessment of feature importances becomes challenging. By conducting thorough analyses, including checking for collinearity through Variance Inflation Factor assessments, researchers can improve the reliability of their findings, thus facilitating more informed conclusions in the application of machine learning across diverse research fields.

Authors' contributions

Yoshiyasu Takefuji completed this research and wrote this article.

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable

Code availability

Not applicable

Funding

This research has no fund.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Mohseni, N., Elhaik, E., 2024. Biases of principal component analysis (PCA) in physical anthropology studies require a reevaluation of evolutionary insights. eLife 13, RP94685. https://doi.org/10.7554/eLife.94685.2.

[2] Qin, W., Zheng, S., Guo, K., Yang, M., Fang, J., 2024. Predicting reaction kinetics of reactive bromine species with organic compounds by machine learning: feature combination and knowledge transfer with reactive chlorine species. J Hazard Mater, 136410. https://doi.org/10.1016/j.jhazmat.2024.136410.

[3] Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20, 177.

[4] Thakur, D., Biswas, S., 2024. Permutation importance based modified guided regularized random forest in human activity recognition with smartphone. Eng Appl Artif Intell 129, 107681. https://doi.org/10.1016/j.engappai.2023.107681.

[5] Sylvain, J.D., Anctil, F., Thiffault, É., 2021. Using bias correction and ensemble modelling for predictive mapping and related uncertainty: a case study in digital soil mapping. Geoderma 403, 115153. https://doi.org/10.1016/j.geoderma.2021.115153.

[6] Chung, H., Park, C., Kang, W.S., Lee, J., 2021. Gender bias in artificial intelligence: severity prediction at an early stage of COVID-19. Published 2021 Nov 29. Front Physiol 12, 778720. https://doi.org/10.3389/fphys.2021.778720.

[7] Bilodeau, B., Jaques, N., Koh, P.W., Kim, B., 2024. Impossibility theorems for feature attribution. Proc Natl Acad Sci USA 121 (2), e2304406120. https://doi.org/10.1073/pnas.2304406120.

[8] Hollman, J.H., Krause, D.A., 2023. Machine learning in admissions?: Use of chi-square automatic interaction detection (CHAID) to predict matriculants to physical therapy school. J Allied Health 52 (3), e93–e98.

[9] Aguilar-Elena, R., Agún-González, J.J., 2024. Chi-square automatic interaction detection (CHAID) analysis of the use of safety goggles and face masks as personal protective equipment (PPE) to protect against occupational biohazards. J Biosaf Biosecurity 6 (2), 125–133. https://doi.org/10.1016/j.jobb.2024.05.001.

[10] Jiang, J., Zhang, X., Yuan, Z., 2024. Feature selection for classification with Spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets. Expert Syst Appl 249 (B), 123633. https://doi.org/10.1016/j.eswa.2024.123633.

[11] Yu, H., Hutson, A.D., 2024. A robust Spearman correlation coefficient permutation test. Commun Stat Theory Methods 53 (6), 2141–2153. https://doi.org/10.1080/03610926.2022.2121144.

[12] Cheng, J., Sun, J., Yao, K., Xu, M., Cao, Y., 2022. A variable selection method based on mutual information and variance inflation factor. Spectrochim Acta A Mol Biomol Spectrosc 268, 120652. https://doi.org/10.1016/j.saa.2021.120652.

[13] Salmerón-Gómez, R., García-García, C.B., García-Pérez, J., 2024. A redefined variance inflation factor: overcoming the limitations of the variance inflation factor. Comput Econ. https://doi.org/10.1007/s10614-024-10575-8.