Journal of Infection 89 (2024) 106357

Contents lists available at ScienceDirect

Journal of Infection

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

Letter to the Editor

Reassessing feature importance biases in machine learning models for infection analysis

Meirman et al. introduced a machine learning model designed to analyze complex infection tendency signals derived from laboratory biomarkers.[1] By employing the XGBoost algorithm with default parameters, they classified individuals within a general cohort as either infected or control. Their study focused on assessing feature importance to identify protective and risk factors associated with infection, utilizing SHAP values to highlight which laboratory biomarkers provided the strongest predictive signals.[1]

Despite the growing popularity of machine learning techniques like XGBoost, it is essential to recognize that the feature importance metrics produced by such models can be misleading due to inherent biases.[2–6] These biases can lead to erroneous conclusions, particularly in specialized fields like infection analysis, where Meirman et al. have considerable expertise but may not fully account for the complexities of algorithmic calculations and their associated biases.

Although the goal of machine learning is to accurately predict outcomes, feature importance metrics reflect model-specific associations rather than genuine relationships. This model-specific nature indicates that different algorithms can produce varying importance metrics, complicating the interpretation of associations. Furthermore, while SHAP values are a valuable tool for interpreting model predictions, they inherently inherit the biases present in the underlying model.

Thus, it is imperative for Meirman et al. to reassess their reliance on potentially biased feature importances generated by the machine learning model. To draw more robust conclusions, they should consider employing rigorous statistical methods to uncover authentic associations between the target variable and its features. Statistical approaches such as Chi-squared tests and Spearman's correlation—both accompanied by p-values—offer bias-free alternatives for validating their findings.[7–10] This paper underscores the intrinsic biases linked to feature importance in machine learning models like XGBoost[2–6] and advocates for stringent statistical analyses[7–10] to enhance the validity and reliability of their results.

Understanding feature importance biases in XGBoost and SHAP necessitates a detailed examination of the methodologies used and the underlying assumptions of these tools. XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm based on decision trees, constructed sequentially to correct errors made by prior trees. While highly effective, the methods for calculating feature importance can introduce significant biases. XGBoost computes feature importance using several metrics, including gain, which measures a feature's contribution based on the accuracy improvement from splits that utilize that feature; cover, which indicates the relative number of observations associated with a feature; and frequency, which counts how often a feature is used across all trees.

The model-specific nature of these metrics highlights that feature importance is contingent on the specific model and its configuration. Thus, altering the model or its parameters can yield different importance rankings, suggesting that these values are not absolute indicators of predictive relevance. Additionally, interactions between features further complicate importance assessments. When features are correlated, the importance assigned to any individual feature can be misleading, as the model may understate or overstate contributions based on its internal allocation of "credit.".

Overfitting presents another critical concern. XGBoost models can overfit to the training data, resulting in inflated feature importance values for features that may not be genuinely predictive in new data. The sequential nature of tree addition can also result in features that correct errors from previous trees receiving disproportionate importance, leading to an exaggerated view of their effectiveness.

While SHAP (SHapley Additive exPlanations) is recognized as a robust interpretation framework, it is not devoid of bias. A key issue is model dependence; SHAP values are conditioned on the training data and the specific structure of the model, which means they can inherit and amplify biases inherent in the model itself. Furthermore, SHAP assumes feature independence when estimating contributions, an assumption that may not always hold true. When features are correlated, their contributions may distort the overall understanding, masking or misrepresenting the effects of individual features.

SHAP values can also be sensitive to outliers or noise in the data, which can lead to misleading assessments of feature importance. Features that are typically insignificant may be unduly impacted by outlier values, resulting in distorted importance metrics. Additionally, in scenarios where features exhibit non-linear relationships with the target variable, SHAP values may misrepresent their contributions, particularly in complex models like XGBoost.

Both XGBoost and SHAP are powerful tools for evaluating feature importance, yet they are not free from biases. The dependencies inherent in XGBoost and the nature of feature interactions can lead to skewed importance measures. Simultaneously, SHAP values can propagate these biases while introducing their own, linked to model dependence and assumptions about feature independence. Given these considerations, researchers should prioritize using rigorous statistical methods to supplement machine learning-derived feature importance, thus providing more reliable insights into the true associations between features and outcomes.

In conclusion, while the machine learning model presented by Meirman et al. utilizes XGBoost to analyze infection tendencies through laboratory biomarkers, the reliance on feature importance metrics raises significant concerns regarding bias and interpretation. The potential for misleading conclusions stemming from model-

https://doi.org/10.1016/j.jinf.2024.106357

specific associations necessitates a critical reevaluation of how feature importance is derived and understood in this context. Although SHAP values offer a structured approach to interpreting model predictions, they are inherently influenced by the model's biases and underlying assumptions, particularly concerning feature independence. To enhance the validity of their findings, it is crucial for the authors to incorporate robust statistical methods—such as Chi-squared tests and Spearman's correlation—into their analysis. By doing so, they can uncover genuine relationships between features and outcomes, ultimately advancing knowledge in infection analysis while mitigating the risks associated with biased interpretations.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

This research has received no funding.

Author contributions

Yoshiyasu Takefuji completed this research and wrote this article.

Data availability

Not applicable.

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.

References

1. Meirman TD, Shapira B, Balicer RD, Rokach L, Dagan N. *Trends of common laboratory biomarkers after SARS-CoV-2 infection. J Infect* 2024;**89**(6):106318. https://doi.org/10.1016/j.jinf.2024.106318
2. Qian H, Wang B, Yuan M, Gao S, Song Y. *Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree. Expert Syst Appl* 2022;**190**:116202. https://doi.org/10.1016/j.eswa.2021.116202
3. 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. https://doi.org/10.1016/j.engappai.2023.107681
4. Demircioğlu A. *Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. Insights Imaging* 2021;**12**(1):172. https://doi.org/10.1186/s13244-021-01115-1
5. Cava W, Bauer C, Moore JH, Pendergrass SA. *Interpretation of machine learning predictions for patient outcomes in electronic health records. AMIA Annu Symp Proc* 2020;**2019**:572–81. Published 2020 Mar 4.
6. Adler AI, Painsky A. *Feature importance in gradient boosting trees with cross-validation feature selection. Entropy* 2022;**24**(5):687. https://doi.org/10.3390/e24050687
7. Asowata OJ, Okekunle AP, Akpa OM, Fakunle AG, Akinyemi JO, Komolafe MA, et al. *Risk assessment score and chi-square automatic interaction detection algorithm for hypertension among Africans: models from the SIREN study. Hypertension* 2023;**80**(12):2581–90. https://doi.org/10.1161/HYPERTENSIONAHA.122.20572
8. Zheng Y, Mao Y, Tsao M, Cowen LLE. *Minimum chi-square method for estimating population size in capture-recapture experiments. PLoS One* 2023;**18**(10):e0292622. https://doi.org/10.1371/journal.pone.0292622
9. Tyagi A, Salhotra R, Agrawal A, Vashist I, Malhotra RK. *Use of Pearson and Spearman correlation testing in Indian anesthesia journals: an audit. J Anaesthesiol Clin Pharmacol* 2023;**39**(4):550–6. https://doi.org/10.4103/joacp.joacp_13_22
10. Jiang J, 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. https://doi.org/10.1016/j.eswa.2024.123633

Yoshiyasu Takefuji [1]
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan
E-mail address: takefuji@keio.jp
[1] **ORCID:** 0000-0002-1826-742X