Journal of Catalysis 446 (2025) 116098

Contents lists available at ScienceDirect

Journal of Catalysis

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

Letter to the Editor

An urgent call for robust statistical methods in reliable feature importance analysis across machine learning

ARTICLE INFO

Keywords:
Machine learning
Feature importance
Pearson correlation
Nonlinear analysis
Statistical methods
Ground truth values

ABSTRACT

Accurate analytical outcomes in machine learning are contingent on error-free calculations and a solid understanding of foundational principles. A notable challenge arises from the lack of ground truth values for validation, complicating the assessment of feature importance, especially when employing linear models with parametric assumptions. This paper critiques the use of Pearson correlation and feature importances derived from Gradient Boosting Regressor (GBR), emphasizing their limitations in analyzing nonlinear and nonparametric data. We propose robust statistical methods, such as Spearman's correlation and Kendall's tau, as alternatives for capturing complex relationships while providing essential directional information. Additionally, attention to Variance Inflation Factor (VIF) is crucial for mitigating feature inflation. By addressing these concerns, researchers can achieve more reliable analyses and deeper insight into variable relationships.

1. Introduction

Accurate analysis relies heavily on error-free calculations, which underscores the importance of researchers being well-versed in the fundamental principles of machine learning. One of the primary challenges in this context is the availability of ground truth values for validation. In supervised machine learning, these ground truth values are essential for verifying target accuracy; however, their presence alone does not guarantee the reliability of derived feature importances, particularly when ground truth data is absent. In such scenarios, meticulous consideration and alternative validation strategies become imperative for ensuring the robustness and accuracy of the analysis. Linear models with parametric assumptions can distort outcomes against nonlinear and nonparametric data.

The terms "nonparametric characteristics," "directional insight," and "model-derived importance scores" are rooted in statistical modeling and data analysis. These concepts aim to extract meaningful information from data, which is increasingly essential when integrating complex experimental datasets with computational approaches.

Nonparametric characteristics refer to statistical methods that do not rely on the assumption of a specific probability distribution for the data, such as a normal or exponential distribution. In many cases, especially in catalysis research, experimental data may exhibit irregular patterns or deviate from standard statistical forms. Nonparametric techniques offer the flexibility needed to analyze such complex data, thus allowing researchers to work effectively with datasets that might challenge traditional parametric methods.

Directional insight involves identifying and interpreting trends or relationships within the data. Rather than considering variables in isolation, this approach examines how changes in one factor may influence another. For example, in catalysis, understanding how variations in reaction conditions—such as temperature, pressure, or catalyst concentration—correlate with changes in reaction rates or product distributions can provide valuable directional insights. Such information not only permits the observation of correlations but also helps to form hypotheses about potential causal relationships that can guide future experimental modifications and optimizations.

Model-derived importance scores are generated by predictive statistical models and provide quantitative measures for ranking or assessing the influence of different variables on a target outcome. In the context of catalysis, these scores help to pinpoint which experimental variables—whether related to reaction conditions, catalyst properties, or process parameters—most significantly affect performance. This quantitative ranking can guide researchers in prioritizing the most impactful factors, leading to more focused and efficient experimental designs.

By understanding these concepts, researchers can bridge the gap between traditional experimental methods and modern data analytics. Integrating nonparametric methods, seeking directional insight, and applying model-derived importance scores enables a more robust and nuanced interpretation of catalytic phenomena. This integrated approach ultimately supports enhanced decision-making and the optimization of experimental designs in catalysis research.

Researchers must have a solid understanding of the fundamental theoretical principles underpinning machine learning, as well as the assumptions associated with the analysis tools they employ. When linear methods are applied to nonlinear data, or when parametric approaches are used on nonparametric data, the results can become distorted or skewed, ultimately leading to erroneous conclusions in the analysis of nonlinear and nonparametric datasets. Therefore, unless the data is confirmed to be linear and parametric, it is imperative for researchers to adopt nonlinear and nonparametric methods.

While supervised machine learning models are beneficial because they have ground truth values for validating target prediction accuracy, the same cannot be said for feature importance metrics derived from these models. The absence of established ground truth values for feature

https://doi.org/10.1016/j.jcat.2025.116098
Received 1 February 2025; Received in revised form 15 March 2025; Accepted 22 March 2025
Available online 24 March 2025

importance can result in significant biases, as different models employ varying methodologies to calculate these metrics. Consequently, this leads to non-negligible discrepancies in feature importance assessments, highlighting the need for careful consideration in their interpretation and application.

Yang et al. investigated stable adsorption configuration searching in hetero-catalysis using methodologies based on similar distribution and active learning [1]. They calculated the Pearson correlation coefficient to assess feature correlations and importance, while feature importances were derived using a Gradient Boosting Regressor (GBR) [1].

This paper highlights two critical concerns regarding the use of the Pearson correlation coefficient and feature importances derived from GBR. The linear and parametric nature of the Pearson correlation coefficient may lead to misleading conclusions when applied to nonlinear and nonparametric data. Yang et al. did not examine linear and parametric nature of data. Additionally, feature importances from machine learning models are inherently biased when ground truth values are absent. Although supervised machine learning methods have access to ground truth values for validating target prediction accuracy, this validation does not guarantee reliable feature importance assessments due to distinct issues inherent in the methodologies used. Various studies of over 100 peer-reviewed articles have documented significant biases in feature importances derived from machine learning models, including GBR, underscoring this challenge [2–7].

Pearson's correlation is a linear, parametric measure that can produce misleading results when applied to nonlinear or nonparametric data, which may lead to inaccurate conclusions. Similarly, feature importance scores generated from Gradient Boosted Regression (GBR) models can be biased due to the absence of ground truth values for validation. To accurately determine the true associations between the target variable and features, three key factors must be taken into account. While SHAP (SHapley Additive exPlanations) is commonly used for feature importance analysis, the function explain=SHAP(model) indicates that SHAP may inherit and potentially amplify the biases present in the model, contributing to incorrect interpretations of feature importance [14–18].

To accurately identify true associations or genuine relationships between a target variable and its features, three critical components must be considered: the data distribution, the statistical relationship between the variables, and the validation of statistical significance through p-values.

This paper advocates for employing nonlinear and nonparametric robust statistical methods, such as Spearman's correlation [8], Kendall's tau [9], Goodman-Kruskal Gamma [10], Somers' D [11], and Hoeffding's D [12], along with their associated p-values. These methodologies excel at capturing complex relationships that may not conform to linear assumptions.

In contrast to feature importance measures, which typically range from 0 to 1, indicating the strength of relationships without directional context, the aforementioned statistical methods provide strength and directional information, with values ranging from −1 to 1. A positive value in these measures indicates a positive association, suggesting that as one variable increases, the other variable also tends to increase. Conversely, a negative value indicates a negative association, meaning that an increase in one variable often corresponds to a decrease in the other. The directional information offered by these statistical methods is invaluable for understanding the nature of relationships among variables. It enables researchers to quantify the strength of associations while also interpreting whether those associations are positive or negative, leading to richer insights into the underlying dynamics of the data. Such comprehensive analyses can facilitate more informed decision-making and hypothesis generation in research.

Furthermore, before applying these statistical methods, it is essential to assess the Variance Inflation Factor (VIF) to identify and eliminate features exhibiting collinearity and interactions [13]. This step is critical for mitigating feature inflation and enhancing the robustness and accuracy of the analysis. By systematically addressing these concerns, researchers can derive deeper insights into the relationships within their data, yielding more reliable conclusions and stronger foundations for inference.

Numerous peer-reviewed articles emphasize that researchers should consider nonlinear behaviors when analyzing catalytic processes [19–22].

Declarations

Funding: This research has no fund.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and material: Not applicable.

Code availability: Not applicable.

Ethics approval: Not applicable.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Writing – review & editing, Writing – original draft, Validation, Investigation, Conceptualization.

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] J. Yang, X. Zhang, X. Zhang, B. Niu, F. Wu, N. Luo, J. He, C. Wang, B. Shan, Q. Li, Stable adsorption configuration searching in hetero-catalysis based on similar distribution and active learning, J Catal. 443 (2025) 115971, https://doi.org/10.1016/j.jcat.2025.115971.

[2] P. Alaimo Di Loro, D. Scacciatelli, G. Tagliaferri, 2-step Gradient Boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy, Stat. Methods Appl. 32 (2023) 237–270 (2023). Doi: 10.1007/s10260-022-00643-4.

[3] P.M. Steiner, Y. Kim, The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases, J. Causal Inference 4 (2) (2016) 20160009, https://doi.org/10.1515/jci-2016-0009.

[4] J. Ugirumurera, E.A. Bensen, J. Severino, J. Sanyal, Addressing bias in bagging and boosting regression models, Sci. Rep.14(1):18452. Published 2024 Aug 8. doi: 10.1038/s41598-024-68907-5.

[5] A. Fisher, C. Rudin, F. Dominici, 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 (2019) 177.

[6] C. Strobl, A.L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinf. 8 (2007) 25, https://doi.org/10.1186/1471-2105-8-25.

[7] T. Salles, L. Rocha, M. Gonçalves, A bias-variance analysis of state-of-the-art random forest text classifiers, Adv. Data Anal. Classification 15 (2021) 379–405, https://doi.org/10.1007/s11634-020-00409-4.

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

[9] S. Chen, A. Ghadami, B.I. Epureanu, Practical guide to using Kendall's τ in the context of forecasting critical transitions, R Soc Open Sci. 9 (7) (2022) 211346, https://doi.org/10.1098/rsos.211346.

[10] J. Metsämuuronen, Directional nature of Goodman–Kruskal gamma and some consequences: identity of Goodman–Kruskal gamma and Somers delta, and their connection to Jonckheere–Terpstra test statistic, Behaviormetrika 48 (2021) 283–307, https://doi.org/10.1007/s41237-021-00138-8.

[11] Y. Li, M. Liang, L. Mao, S. Wang, Robust estimation and variable selection for the accelerated failure time model, Stat Med. 40 (20) (2021) 4473–4491, https://doi.org/10.1002/sim.9042.

[12] A.S. Eisele, M. Tarbier, A.A. Dormann, V. Pelechano, D.M. Suter, Gene-expression memory-based prediction of cell lineages from scRNA-seq datasets, Nat Commun. 15 (1) (2024) 2744, https://doi.org/10.1038/s41467-024-47158-y.

[13] J. Jacob, R. Varadharajan, Robust variance inflation factor: a promising approach for collinearity diagnostics in the presence of outliers, Sankhya B 86 (2024) 845–871, https://doi.org/10.1007/s13571-024-00342-y.

[14] B. Bilodeau, N. Jaques, P.W. Koh, B. Kim, Impossibility theorems for feature attribution, Proc Natl Acad Sci USA 121 (2) (2024) e2304406120, https://doi.org/10.1073/pnas.2304406120.
[15] X. Huang, J. Marques-Silva, On the failings of Shapley values for explainability, Int J Approx Reason 171 (2024) 109112, https://doi.org/10.1016/j.ijar.2023.109112.
[16] M.A. Lones, Avoiding common machine learning pitfalls, Patterns 5(10) (2024) 101046, doi: 10.1016/j.patter.2024.101046.
[17] C. Molnar, et al., General pitfalls of model-agnostic interpretation methods for machine learning models. In: A. Holzinger, R. Goebel, R. Fong, T. Moon, K.R. Müller, W. Samek, eds. xxAI - Beyond Explainable AI. xxAI 2020. Vol 13200. Lecture Notes in Computer Science. Springer, 2022. doi:10.1007/978-3-031-04083-2_4.
[18] I. Kumar, C. Scheidegger, S. Venkatasubramanian, S. Friedler, Shapley residuals: quantifying the limits of the shapley value for explanations, Adv. Neural Inf. Proces. Syst. 34 (2021) 26598–26608.
[19] L.C. Mayer, S. Heitsch, O. Trapp, Nonlinear effects in asymmetric catalysis by design: concept, synthesis, and applications, Acc Chem Res. 55 (23) (2022) 3345–3361, https://doi.org/10.1021/acs.accounts.2c00557.
[20] C. Ali, D.G. Blackmond, J. Burés, Kinetic rationalization of nonlinear effects in asymmetric catalytic cascade reactions under Curtin–Hammett conditions, ACS Catal. 12 (10) (2022) 5776–5785, https://doi.org/10.1021/acscatal.2c00783.
[21] R. Baratti, J. Alvarez, M. Morbidelli, Design and experimental verification of a nonlinear catalytic reactor estimator, Chem. Eng. Sci. 48 (14) (1993) 2573–2585, https://doi.org/10.1016/0009-2509(93)80268-U.
[22] J.R. Kittrell, W.G. Hunter, C.C. Watson, Obtaining precise parameter estimates for nonlinear catalytic rate models, AIChE J. 12 (1) (1966) 5–10, https://doi.org/10.1002/aic.690120104.

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.