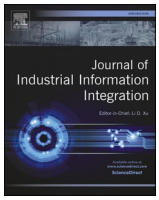
Journal of Industrial Information Integration 45 (2025) 100809

Contents lists available at ScienceDirect

Journal of Industrial Information Integration

journal homepage: www.sciencedirect.com/journal/journal-of-industrial-information-integration



Challenges in feature importance interpretation: Analyzing LSTM-NN predictions in battery material flotation

Yoshiyasu Takefuji *

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

ARTICLE INFO

Keywords
LSTM-NN
Feature importance
Spearman correlation
SAGE
SHAP
Machine learning

ABSTRACT

Gomez-Flores et al. proposed a Long Short-Term Memory Neural Network (LSTM-NN) for predicting the flotation behavior of battery active materials using various physicochemical and hydrodynamic variables. While they achieved high prediction accuracy, validated through Mean Relative Error (MRE) and Mean Squared Error (MSE) metrics, concerns arise regarding the integrity of feature importance assessments derived from SAGE and SHAP methodologies. Specifically, the reliance on these model-specific techniques can introduce biases, obscuring the true relationships between features. Additionally, while Spearman's correlation elucidated significant relationships among variables, the absence of discussion on p-values left gaps in interpretation. This study emphasizes the need for cautious interpretation of feature importance metrics and the elimination of less significant variables, aiming to enhance model robustness and improve actionable insights in machine learning contexts.

Fundamental principles of machine learning aimed at novices and researchers in the field of industrial integrated information. This addition seeks to clarify common misunderstandings surrounding machine learning concepts and practices. The primary goal of machine learning is to accurately predict a target variable using ground truth values to validate target accuracy. In contrast, feature importances from machine learning models are intended to capture the associations between the target and features, even when ground truth values are not present. This absence of definitive ground truth in feature importance calculations means that different models utilize diverse methodologies, leading to varying feature importance values. Such variability can result in biased interpretations and potentially erroneous conclusions.

Moreover, feature importances derived from machine learning models are inherently influenced by the specific characteristics of each model, introducing additional biases. While there are several bias mitigation techniques available, none can fully eliminate these biases in feature importance assessments. Therefore, it is crucial for researchers and practitioners to be aware of these limitations and to carefully interpret feature importances within the context of the chosen modeling approach. A critical evaluation of feature importances is necessary to ensure robust decision-making and to enhance the overall reliability of machine learning outcomes.

To accurately determine the true associations between the target and features, three key elements must be considered: data distribution, the examination of relationships between variables, and statistical validation through p-values. When analyzing data distribution, choosing between linear or nonlinear models, as well as parametric or nonparametric methods, is crucial for obtaining reliable insights.

Robust, bias-free statistical methods, such as Spearman's correlation and Kendall's tau, which are both nonlinear and nonparametric, provide valuable alternatives for assessing feature importance. Failure to appropriately select a model—whether linear or parametric—in the context of inherently nonlinear or nonparametric data can lead to significant distortions in data analysis and ultimately result in misleading conclusions.

Thus, ensuring a thorough understanding of these foundational concepts and carefully evaluating the appropriate methodologies are essential for deriving accurate insights in machine learning applications. This meticulous approach will enhance the robustness of conclusions drawn from data analyses and support the ongoing development of reliable machine learning models.

Many researchers often misinterpret the fundamental principles of machine learning. For instance, while cross-validation techniques that manipulate data shuffles are effective for validating target prediction accuracy, they do not necessarily ensure the accuracy of feature importance assessments. Similarly, diverse metrics such as R-squared and RMSE, which measure target prediction accuracy, do not guarantee valid feature importance rankings. This misunderstanding can lead to

* Corresponding author at: Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan.
E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.jii.2025.100809

Available online 24 February 2025

the selection of misleading features or the misinterpretation of their impact on the target variable, ultimately compromising the integrity of the modeling process. It is essential to recognize that while target accuracy metrics provide valuable insights into model performance, they should not be conflated with metrics assessing the accuracy of feature importance. Understanding these distinctions is crucial for robust machine learning practices.

With the advent of explainable AI, SHAP (SHapley Additive exPlanations) has gained significant popularity in research for its ability to provide insights into model predictions. However, one key limitation of SHAP is that its function, explain=SHAP(model), inherits biases from the underlying model. This can lead to the amplification of these biases, resulting in potentially misleading explanations of model behavior.

In contrast, statistical measures such as Spearman's correlation and Kendall's tau offer model-agnostic approaches that do not depend on the specific characteristics of a machine learning algorithm. By focusing on the rank-order relationships between variables, these methods provide a more unbiased perspective on feature associations.

This paper advocates for the use of bias-free, robust statistical methods instead of relying on SHAP, which may utilize biased feature importances. Emphasizing model-less approaches enhances the credibility of interpretations and allows for more reliable insights into the relationships between features and the target variable. By prioritizing unbiased statistical methods, researchers can foster a clearer understanding of the data and avoid the pitfalls associated with biased model interpretations.

Gomez-Flores et al. introduced a Long Short-Term Memory Neural Network (LSTM-NN) for predicting the flotation behavior of battery active materials based on a range of physicochemical and hydrodynamic variables [1]. To enhance the predictive accuracy of their model, they calculated Spearman's correlation coefficients for the dataset employed in their machine learning analysis. Furthermore, they utilized the Shapley Additive Global Importance (SAGE) methodology to evaluate the significance of input variables within the model, with additional details available in related publications [1].

While this study acknowledges the impressive predictive performance of the LSTM-NN, as evidenced by robust Spearman correlation values, it raises critical concerns regarding the interpretability of feature importance derived from SAGE and SHAP methods. The intrinsic model-specific nature of these interpretive techniques can lead to misleading conclusions regarding the true significance of features. Over 100 peer-reviewed articles have explored biases in feature importance and selection derived from machine learning models. Notably, existing models, including Long Short-Term Memory Neural Networks (LSTM-NN), are prone to inducing inherent biases in feature importance assessments [2–5]. These biases can significantly compromise the integrity and reliability of the resulting interpretations, leading to potentially misleading conclusions about the significance of various features in predictive analyses. Addressing these biases is crucial for enhancing the validity of machine learning applications and ensuring that decisions based on model outputs are well-informed.

Moreover, although Gomez-Flores et al. demonstrated high prediction accuracy and validated their model using Mean Relative Error (MRE) and Mean Squared Error (MSE) metrics, these validation measures do not inherently validate the accuracy of the corresponding feature importance assessments. While they successfully established genuine correlations between the target variable and input features through Spearman's correlation—augmented by statistical p-values—these p-values were not explicitly discussed in their findings. This oversight leaves a significant gap in the interpretation of their results.

When analyzed in conjunction with Spearman's correlation, p-values can effectively determine whether the observed feature importances are statistically significant or merely the result of random chance. This dual analysis, therefore, significantly enhances the robustness of the findings. It is crucial to recognize that features with higher p-values may not be relevant and should be considered for elimination from the final selection of input variables. In doing so, one can strengthen the model's interpretability and ensure that only statistically significant features influence the predictions, ultimately leading to more reliable and actionable insights.

The reliance on SAGE and SHAP methodologies poses challenges, as they are closely tied to the specific machine learning model employed, which can amplify any inherent biases present in that model. The function explain=SHAP(model) not only reflects but also perpetuates these biases, leading to discrepancies between the Spearman correlation values and the feature importance metrics produced by SAGE and SHAP. This discrepancy underscores the limitations of these interpretive methods, raising valid concerns regarding the reliability of feature importance assessments generated through these techniques. As a result, these methods may fail to accurately portray the true influences of input variables on the model's predictions, potentially misguiding subsequent analyses and applications.

Currently, there is no comprehensive tool available to entirely mitigate biases in feature importance assessments derived from machine learning models or to precisely quantify these biases. The absence of a robust solution highlights the pressing need for caution when interpreting feature importance metrics from complex models. Critical evaluation and further research are essential to enhance the reliability and fidelity of such assessments in machine learning contexts, ultimately advancing our understanding of feature significance and improving model interpretability.

Funding

This research has no fund.

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

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] A. Gomez-Flores, H. Park, G. Hong, H. Nam, J. Gomez-Flores, S. Kang, G.W. Heyes, L de S Leal Filho, H. Kim, J.M. Lee, J. Lee, Flotation separation of lithium–ion battery electrodes predicted by a long short-term memory network using data from

physicochemical kinetic simulations and experiments, J. Ind. Inf. Integr. 42 (2024) 100697, https://doi.org/10.1016/j.jii.2024.100697.
[2] 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.
[3] L. Ren, T. Wang, Y. Laili, L. Zhang, A data-driven self-supervised LSTM-DeepFM model for industrial soft sensor, IEEE Trans. Ind. Inf. 18 (9) (2022) 5859–5869, https://doi.org/10.1109/TII.2021.3131471.
[4] M. Moran, G. Gordon, Deep curious feature selection: a recurrent, intrinsic-reward reinforcement learning approach to feature selection, IEEE Trans. Artif. Intell. 5 (03) (2024) 1174–1184, https://doi.org/10.1109/TAI.2023.3282564.
[5] P. Takaew, J.C. Xia, L.S. Doucet, Machine learning and tectonic setting determination: bridging the gap between Earth scientists and data scientists, Geosci. Frontiers 15 (1) (2024) 101726, https://doi.org/10.1016/j.gsf.2023.101726.