

Explainable XGBoost Prediction of Employee Turnover Intention Using Organizational Behavior and Human Resource Features

Selin. Karaca¹, Julien. Moreau², Youssef. El Amrani³

¹ Department of Business Administration and Management, Middle East Technical University, Ankara, Turkey

² Department of Organizational Behavior and Human Resources, HEC Paris, Paris, France

³ Department of Organizational Management, Mohammed V University, Rabat, Morocco

* Corresponding author email address: julien.moreau@hec.fr

E d i t o r	R e v i e w e r s
Aliakbar Aminbeidokhti ^{id} Educational Administration, Faculty of Psychology and Educational Sciences, Semnan University, Semnan, Iran kafashpor@um.ac.ir	Reviewer 1: Mehrdad Bayat ^{id} Assistant Professor, Department of Management, Payam Noor University, Tehran, Iran. Email: bayatmehrdad60@pnu.ac.ir Reviewer 2: Manijeh Haghighinasab ^{id} Assistant Professor, Department of Management, Alzahra University, Tehran, Iran Email: haghighinasab@srbiau.ac.ir

1. Round 1

1.1. Reviewer 1

Reviewer:

In the Study Design and Participants paragraph, the sentence “employees who were self-employed, temporary interns, or had less than six months of organizational tenure were excluded” requires stronger conceptual justification. Excluding employees with short tenure may systematically remove individuals at particularly high risk of early turnover and therefore alter the observed outcome distribution. This is especially important because organizational tenure later appears as a predictor in the model. The authors should explain why six months was selected as the minimum tenure threshold, whether this criterion was established before data collection, and how excluding newly hired employees might affect the generalizability of the model to early-stage employee turnover.

In the Data Collection Tools section, the manuscript states that organizational behavior constructs were measured using “multi-item Likert-type statements adapted from established organizational behavior and human resource measurement approaches,” but it does not identify the exact instruments, original authors, item numbers, scoring procedures, or psychometric

evidence. This is a major reproducibility concern. Constructs such as job satisfaction, organizational commitment, perceived organizational support, work engagement, psychological safety, organizational justice, and person–organization fit have multiple validated measures with different theoretical foundations. The authors should name each scale, cite its source, report the number of items, response range, score construction, French translation procedure where applicable, and reliability indices observed in the present sample.

In the Data Collection Tools paragraph, the operationalization of turnover intention requires substantial clarification. The manuscript states that “responses were combined into an overall turnover-intention score” and subsequently dichotomized into high and low categories, but the precise number of items, scoring algorithm, and cutoff are not reported. More importantly, converting a continuous or ordinal turnover-intention scale into a binary outcome can reduce information and may create artificial distinctions between employees whose scores are very similar. The manuscript should report the exact threshold, justify it theoretically or empirically, and provide a sensitivity analysis showing whether model performance and feature importance remain stable under alternative thresholds or when turnover intention is modeled as a continuous outcome.

The Methods section should address measurement validation before model development. The current text notes only that “internal consistency was evaluated for each multi-item construct before inclusion in the predictive analysis,” but no reliability or construct-validity results are reported. At minimum, Cronbach’s alpha and preferably McDonald’s omega should be provided for multi-item scales. Because several constructs overlap conceptually, the authors should additionally assess discriminant validity and redundancy. If composite scores rather than individual questionnaire items were used as predictors, confirmatory factor analysis or another appropriate measurement-validation procedure would considerably strengthen confidence that the predictors represent empirically distinct constructs.

In Table 3, the reported SHAP ranking should be interpreted with greater methodological caution. Mean absolute SHAP values quantify average contribution magnitude within the fitted model; they do not establish causal importance, substantive priority, or independent effects. The sentence stating that job satisfaction was “the most influential predictor” is acceptable if clearly framed as model-specific predictive importance, but subsequent interpretations occasionally move toward causal language. The authors should consistently distinguish predictive contribution from causal effect and explicitly acknowledge that correlated predictors can share, redistribute, or mask SHAP importance.

The SHAP analysis requires additional treatment of correlated features. Several reported variables—including job satisfaction, organizational commitment, perceived organizational support, leadership quality, supervisor support, recognition, HR fairness, and work engagement—are likely to be substantially correlated. SHAP values can behave differently depending on assumptions about feature dependence, and correlated predictors may lead to unstable allocations of importance. The manuscript should report the correlation structure among the principal features, explain which SHAP computation approach was used, and preferably conduct grouped or conditional SHAP analyses to determine whether the substantive ranking remains stable when conceptually related predictors are considered together.

Authors revised the manuscript and uploaded the new document.

1.2. Reviewer 2

Reviewer:

In the Data Analysis paragraph, the authors state that the dataset was divided into 80% training and 20% independent testing samples, yielding 547 and 137 observations, respectively. A test set of only 137 cases, of which approximately 50 would belong to the high-turnover class, may yield considerable uncertainty around performance estimates. The manuscript should therefore report 95% confidence intervals for ROC-AUC, accuracy, sensitivity, specificity, precision, and F1 score. Bootstrap confidence intervals would be appropriate. Without uncertainty estimates, differences such as an ROC-AUC of 0.91 during cross-validation and 0.88 in testing cannot be meaningfully interpreted as evidence of stable generalization.

In the Data Analysis section, hyperparameter optimization is insufficiently reproducible. The manuscript states that “the number of boosting trees, learning rate, maximum tree depth, minimum child weight, subsampling ratio, column-sampling

ratio, and L1 and L2 regularization parameters were optimized,” but neither the search strategy nor the final hyperparameter values are reported. The authors should specify whether tuning used grid search, randomized search, Bayesian optimization, or another procedure; give the search ranges; identify the optimization metric; state the random seed; and report the selected values. This information is essential for computational reproducibility and for determining whether model complexity was appropriately controlled.

The manuscript should strengthen protection against data leakage. Although it appropriately states that preprocessing was conducted using the training data only, it is unclear whether hyperparameter tuning, missing-value imputation, categorical encoding, feature filtering, class weighting, and any feature-selection procedures were all nested within each cross-validation fold. The sentence stating that “all preprocessing procedures...were conducted using information from the training data only” is not sufficiently precise. The authors should explicitly describe a unified modeling pipeline in which every data-dependent transformation was fitted independently within the training portion of each fold to prevent optimistic cross-validation estimates.

In the Findings paragraph reporting model performance, the interpretation that “the model generalized reasonably well to previously unseen observations and did not exhibit pronounced overfitting” is somewhat stronger than the reported evidence permits. A modest reduction in ROC-AUC from 0.91 to 0.88 is encouraging, but overfitting cannot be excluded solely from this difference, particularly with one relatively small holdout sample. The wording should be tempered, and the authors should supplement the evaluation with repeated nested cross-validation, bootstrapped optimism correction, or external validation. The distinction between cross-validation performance and truly external generalization should be made explicit.

In Table 2 and its accompanying paragraph, predictive performance is reported without comparison against any benchmark or alternative model. Demonstrating that XGBoost achieves an ROC-AUC of 0.88 does not establish that XGBoost meaningfully improves prediction compared with simpler, more interpretable models. The authors should compare XGBoost with at least logistic regression and preferably random forest, support vector machine, or another relevant classifier using identical resampling procedures. Statistical comparison of ROC-AUC values or paired resampling differences would allow readers to determine whether the complexity of XGBoost produces a substantive predictive advantage.

The model evaluation framework should place greater emphasis on calibration. Although Brier score and log loss are reported, the manuscript does not present a calibration curve, calibration intercept, calibration slope, or observed-versus-predicted probabilities. This is important because a model can discriminate well while systematically overestimating or underestimating turnover risk. Since the paper frames the model as potentially useful for HR decision support, calibration is arguably as important as discrimination. A calibration plot and quantitative calibration measures should therefore be added, particularly if predicted probabilities are intended to identify employees or groups at elevated risk.

Authors revised the manuscript and uploaded the new document.

2. Revised

Editor’s decision after revisions: Accepted.

Editor in Chief’s decision: Accepted.