Journal of Organizational and End User Computing
Volume 38 • Issue 1 • January-December 2026

Employee Turnover Prediction Research of Human Resource Management on Machine Learning Algorithms and Big Data Analysis

Rongjie Qin
https://orcid.org/0009-0004-7686-3344
Wuhan Technology and Business University, China & Research Center for Hubei Business, Service, and Development, China

Xiaolin Qi
https://orcid.org/0009-0009-6407-773X
Wuhan Technology and Business University, China & Research Center for Hubei Business, Service, and Development, China

Ying Yuan
Wuhan Technology and Business University, China & Research Center for Hubei Business, Service, and Development, China

Bilal Alatas
https://orcid.org/0000-0002-3513-0329
Firat University, Turkey
Received: September 20th, 2025 | **Accepted:** January 2nd, 2026

ABSTRACT

This study introduces a new tool for predicting employee turnover using machine learning (ML) and big data. This method integrates LightGBM and XGBoost (both weighted 1, with predictions summed) to enhance accuracy and stability. To improve model interpretability, the SHAPT model is used to identify key factors affecting turnover, such as salary, position, and tenure. Experimental results show the integrated model outperforms standalone LightGBM and XGBoost: accuracy is 1.5% higher, F1 value is 0.02 higher, and AUC reaches 0.9504. These validate the model; SHAP analysis also provides actionable HR management insights, enabling early identification and response to potential employee departures. The research offers practical tools for HR decision-making. Future work will incorporate additional socio-economic variables and dynamic data to further improve prediction performance.

KEYWORDS

Employee Turnover Prediction, Machine Learning, Weighted Sum, SHAP Model, Big Data Analysis, Interpretable Analysis

INTRODUCTION

In the contemporary competitive market, employee turnover has become an essential problem of enterprise human resources (HR) management (Islam et al., 2023). Employee turnover not only will cause operating costs to increase, but it will also impact the stability of the team and the continuation of corporate culture. Thus, accurately predicting the turnover risk of employees and taking measures to intervene effectively is crucial for improving the competitiveness of enterprises (Chowdhury et al.,

DOI: 10.4018/JOEUC.399146

2023). With rapidly evolving big data technologies and machine learning (ML) algorithms, it is possible to employ data-driven approaches to predict employee turnover to not only provide forward-looking assistance to enterprises in making decisions, but also help improve HR management and mitigate adverse effects of employee turnover (Yahia et al., 2021).

However, existing employee turnover prediction methods still have significant limitations, especially in terms of interpretability, where challenges remain unresolved. In terms of traditional statistical methods, models such as logistic regression and decision trees can handle some simple factors affecting turnover (such as the correlation between salary level and turnover rate in a single dimension), but they have two major shortcomings. First, they are unable to capture the complex nonlinear relationships and interaction effects between features (such as the joint impact of "years of service + job level" on turnover decisions), and their prediction accuracy drops significantly on high-dimensional, large-scale datasets (Ning et al., 2025; Pourkhodabakhsh et al., 2023). Second, their interpretability is limited to the surface level of whether features are related, and they cannot quantify the degree of influence of features on turnover outcomes—for example, the coefficients of logistic regression can only reflect the directional correlation of features but cannot explain "how much the risk of turnover increases when an employee's monthly income is 10% lower than the average;" they also cannot distinguish the differentiated impact of features on different individuals, making it difficult for companies to formulate targeted intervention measures based on the prediction results (Pessach et al., 2020; Qi et al., 2022).

In recent years, ML techniques (such as random forests, support vector machines, and neural networks) have been widely used in turnover prediction and have performed well in identifying complex patterns and hidden associations in data (Folorunso et al., 2021; Pratt et al., 2021). However, the contradiction between "high accuracy and low interpretability" has become more prominent. These models are mostly "black box" models: random forests improve accuracy by ensemble multiple decision trees, but their output depends on a complex voting mechanism between trees, making it impossible to trace the formation logic of individual prediction results; support vector machines handle nonlinear problems through kernel mapping, but they are unable to explain the correlation between features and turnover results in high-dimensional space (Biason, 2020; Zhao et al., 2024). Although some studies have attempted to introduce interpretive tools such as Shapley additive explanations (SHAP) to identify the features that have the greatest impact on turnover (Masenya, 2024; Mosca et al., 2022; Quan & Lu, 2024), current practices still have shortcomings: most studies only combine SHAP with a single model, failing to address the interpretability problem of ensemble models (such as multi-gradient boosting model fusion). At the same time, existing ensemble models often suffer from structural complexity, long training time, and low application flexibility—for example, some deep ensemble models require a large amount of computing resources, and the model parameter adjustments are not adaptable to different industry datasets, making it difficult to meet enterprises' actual needs for "high accuracy + interpretability + lightweight" (J. Li et al., 2025; P. Li et al., 2024).

To deal with the above challenges, this paper introduces a new weighted sum model—an integrated LightGBM and XGBoost model combined with SHAP—in order to improve the accuracy and interpretability of the prediction (Yao et al., 2022). This method systematically integrates the advantages of two gradient boosting models through a weighted summation strategy—LightGBM's efficiency and nonlinear fitting capabilities in large-scale data processing, and XGBoost's strong stability brought by its regularization mechanism—effectively avoiding prediction biases of single models in complex scenarios. Simultaneously, the SHAP framework is used to perform quantitative attribution analysis on core features, breaking down the "black box" barrier of traditional ensemble models and providing enterprises with accurate evidence for analyzing the driving factors of employee turnover. This paper aims to provide enterprises with a scientifically sound and practical solution for HR decision-making through a collaborative design of "high-precision prediction + interpretability empowerment."

The main contributions of this paper are as follows: