RESEARCH ARTICLE OPEN ACCESS

An End-to-End Framework for Predicting Employee Attrition with Explainable AI

Swaroop Bharat Daund
COEP Technological University, Shivajinagar, Pune-411005
daundsb23.comp@coeptech.ac.in

Abstract:

Employee attrition is a persistent challenge for organizations, generating significant costs and impacting productivity. While machine learning can predict employee turnover, common approaches struggle with imbalanced data and lack interpretability for actionable insights. This paper presents a complete, end-to-end framework for attrition prediction that is both accurate and transparent. Using the IBM HR Analytics dataset, the methodology features advanced preprocessing, engineered attributes, and class balancing via the Synthetic Minority Oversampling Technique (SMOTE). A Random Forest classifier is employed, with model evaluation based on Recall, F1-Score, Precision, and AUC-ROC — metrics aligned with the business goal of accurately identifying at-risk employees. To ensure decision transparency, Explainable AI tools SHAP and LIME are integrated, providing both global and local interpretability. Results reveal that the proposed model identifies 80% of actual attrition cases while highlighting key drivers such as overtime workload, compensation levels, and job satisfaction. This interpretability supports targeted retention strategies instead of generic interventions. The study demonstrates that combining machine learning with XAI techniques enables both reliable prediction and actionable insights. Future work includes integrating career progression timelines and deploying real-time attrition monitoring dashboards.

1. Introduction

Employee attrition — whether voluntary or involuntary — continues to be a major concern for businesses due to its ripple effects on costs, productivity, and team stability. Besides recruitment expenses, organizations face losses in institutional expertise and workforce efficiency. This has driven companies toward proactive retention strategies rather than reactive measures.

Machine learning solutions have increasingly been explored for attrition forecasting, yet conventional approaches face two main challenges:

1. Imbalanced Data – The number of employees who remain far exceeds those who resign, often leading models to favor the majority class and overlook at-risk employees.
2. Limited Interpretability – Many predictive algorithms function as opaque “black boxes,” offering predictions without clarity on why decisions are made, thus reducing trust and actionability.

In this work, we address both issues by balancing the training dataset using SMOTE, employing business-aligned performance metrics, and integrating SHAP and LIME to make predictions interpretable and actionable.

2. Materials and Methods

2.1 Dataset Description

The experiments were conducted using the IBM HR Analytics Employee Attrition & Performance dataset, containing 1,470 records with 35 attributes covering demographic, job-related, and satisfaction features. The target variable, *Attrition*, is binary: “Yes” (16.1%) vs. “No” (83.9%).

2.2 Data Preparation and Feature Engineering

No missing values were detected in the dataset. The following attributes were removed due to zero variance or irrelevance:

- Zero variance: EmployeeCount, Over18, StandardHours
- Unique identifiers: EmployeeNumber

To improve model predictive power, custom features were engineered as follows:

Feature Name	Formula	Rationale
Tenure To Age Ratio	Years At Company / Age	Measures Career Velocity; Low Value Suggests Stagnation
Income Per Year of Service	Monthly Income / (Total Working Years + 1)	Evaluates Compensation Growth; Low Value May Indicate Dissatisfaction
Years Without Promotion	Years At Company – Years Since Last Promotion	Quantifies Career Stagnation
Satisfaction Income Interaction	Job Satisfaction ×Monthly Income	Captures Combined Effect of Satisfaction and Pay

Categorical attributes were converted into numerical codes using Label Encoding.

2.3 Dealing with Class Imbalance

A 70–30 train–test split was applied, and SMOTE was used exclusively on the training set to synthetically balance the classes and mitigate bias toward the majority class.

2.4 Model Selection and Metrics

A Random Forest Classifier was selected for its capability to handle complex, nonlinear patterns and high-dimensional data.

Performance metrics included:

- Recall (priority to reduce false negatives)
- Precision
- F1-Score
- AUC-ROC

2.5 Explainable AI Tools

Two leading interpretability frameworks were employed:

- SHAP: For global feature ranking and per-instance explanations using Shapley values.
- LIME: For locally interpretable results, explaining why a specific prediction was made.

3. Results and Discussion

3.1 Classification Performance

The SMOTE-balanced RF model showed the following results:

Class	Precision	Recall	F1-Score	Support
No (Stay)	0.95	0.92	0.94	370
Yes (Leave)	0.71	0.80	0.75	71
Accuracy			0.90	441
Macro Avg	0.83	0.86	0.84	441
Weighted Avg	0.91	0.90	0.90	441

The recall score of 0.80 for the attrition class indicates the model successfully identified 80% of employees at risk, meeting a crucial business requirement. The AUC-ROC score of 0.92 confirms high separability between the two classes.

3.2 Interpretability with SHAP and LIME

SHAP analysis revealed that *OverTime*, *MonthlyIncome*, *JobSatisfaction*, and *TenureToAgeRatio* were the most influential predictors — consistent with HR research emphasizing work-life balance, compensation, and career progression.

- High-risk indicators: High overtime workload, low satisfaction, below-average income
- Low-risk indicators: Competitive compensation and higher job satisfaction

Such interpretability allows HR teams to develop personalized retention policies instead of generic initiatives.

4. Conclusion

This study proposes a comprehensive, interpretable approach for employee attrition prediction. By combining SMOTE, a Random Forest classifier, and business-relevant evaluation metrics with SHAP and LIME explanations, the framework not only forecasts with high accuracy but also clarifies the *reasons* behind predictions. This transparency enables HR departments to make data-driven retention decisions.
Future work will focus on:

- Integrating temporal patterns of employee career growth
- Developing real-time attrition tracking dashboards
- Deploying the solution as a web-based HR analytics application

Acknowledgments

The authors thank the developers of the IBM HR Analytics dataset for making the data publicly available. No external funding or grants were received for this study.

References

(APA 7th Edition Style)
Alqahtani, H., Almagrabi, H., & Alharbi, A. (2024). Employee attrition prediction using machine learning models: A review paper. *International Journal of Artificial Intelligence and Applications*.
Journal of Computer System and Informatics (JoSYC). (2022). Employee attrition prediction using feature selection with information gain and random forest classification, *3*(4), 410–419.
Mansor, N., Sani, N. S., & Aliff, M. (2021). Machine learning for predicting employee attrition. *International Journal of Advanced Computer Science and Applications*, *12*(11).

Raza, A., et al. (2022). Predicting employee attrition using machine learning approaches. *Applied Sciences, 12*(13), 6424.

Tang, Z., Gu, J., & Kelkar, M. (2025). Enhancing employee retention: Predicting attrition using machine learning models. *Journal of Applied Business and Economics, 27*(3).

ResearchGate. (2018). Employee Attrition Prediction.

ResearchGate. (2020). A comparative study on machine learning algorithms for employee attrition prediction.

ResearchGate. (2021, March). Prediction of employee attrition using machine learning and ensemble methods.

ResearchGate. (2021, August). Prediction of employee attrition using machine learning approach.