Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 10, 2025

Explainable Artificial Intelligence for Employee Attrition Prediction: Integrating Predictive Performance with Actionable HR Retention Strategies

Javed Iqbal Bangash[1], Sajid Rahman Khattak[2] and Majid Khan[3]

[1] Institute of Computer Sciences & Information Technology (ICS/IT), The University of Agriculture, Peshawar – Pakistan.

[2] Institute of Business & Management Sciences (IBMS), The University of Agriculture, Peshawar – Pakistan. ***(Corresponding Author)***

[3] Department of Mathematics, Statistics and Computer Science, The University of Agriculture, Peshawar – Pakistan.

javed.bangash@aup.edu.pk[1], srkhattak@aup.edu.pk[2], majidkhan@aup.edu.pk[3]

DOI: https://doi.org/10.5281/zenodo.22672830

Keywords
Employee Attrition; HR Analytics; Machine Learning; Explainable AI; Permutation Importance; People Analytics; Feature Importance; Employee Retention

Article History
Received on 23 Sept 2025
Accepted on 21 Oct 2025
Published on 30 Oct 2025

Corresponding Author:
Sajid Rahman Khattak
Institute of Business & Management Sciences (IBMS), The University of Agriculture, Peshawar – Pakistan.
srkhattak@aup.edu.pk

Abstract

Employee attrition is one of the most critical phenomena for contemporary companies as it directly impacts the bottom line. Although the application of machine learning (ML) methods for attrition prediction is a widespread practice, the task of providing these black-box models with any practical interpretability for end-users (e.g., human resources (HR) managers) remains underexplored. This paper aims to design an integrated framework that (i) identifies the best-performing ML classifiers for the given task and (ii) employs model-agnostic techniques for feature importance explanation. An extensive comparative performance analysis was carried out on the IBM HR Analytics Employee Attrition Dataset (N = 1470, 31 retained features). Six supervised classification tasks were implemented and evaluated using an 80/20 stratified train/test split and 5-fold stratified cross-validation. The best overall performance was reported by the gradient-boosted tree ensemble (held-out ROC-AUC = 0.812, accuracy = 86.1%), but class-weighted logistic regression demonstrated the best stability across cross-validation folds (mean ROC-AUC = 0.827) and significantly better minority-class recall (61.7%). In line with permutation importance analysis, partial dependence plots, and tree-specific interpretation, logistic regression coefficients identified the vital predictors of employee attrition, namely, work-life balance, stock options, satisfaction, time since the last promotion, business travel, and job role. The patterns discovered were incorporated into a conceptual framework, which included five specific actions for human resource managers to consider in order to improve employee retention. The findings were evaluated through the lens of evidence-based business practices, thus, demonstrating how the combined use of ML and XAI approaches could actualize a predictive model in practice.

INTRODUCTION

Voluntary employee turnover stands as one of the most expensive and productivity-damaging phenomena for modern enterprises. Despite the costs of personnel replacement being substantial (often ranging from half an employee's salary to two times that), attrition incurs additional losses due to the loss of knowledge and experience. Moreover, frequent or mass voluntary resignations could trigger a wave of similar actions, thus provoking a cascade of exits. With the increasing competitiveness of the labor market and the availability of remote work, HR departments face growing challenges in retaining their staff. They are turning to people analytics and predictive modeling to identify employees at highest risk of intent to leave before such events happen, so that appropriate remedial actions could be undertaken.

Machine learning classifiers, ranging from simple linear models to modern ensemble and neural network approaches, have been repeatedly proven efficient at predicting attrition in research articles. However, while being generally accurate, such models typically lack interpretability, providing little to no information on why a particular instance was predicted as a positive (i.e., likely to quit). For an HR business partner, this could be an issue, as most if not all suggested interventions would have little basis or would require additional analysis to answer why questions. Meanwhile, HR departments are legally bound to ensure that their actions do not discriminate against protected classes of employees. Thus, the combination of domain-specific needs and the general desire for model transparency has led to the birth of the field of explainable AI. It aims to provide a means of obtaining human-interpretable explanations of why a black-box model made a prediction it did, without sacrificing the accuracy.

The current research endeavors to contribute to the goal by providing a fully descriptive procedure of analyzing the attrition phenomenon. First, the performance of six different machine learning algorithms is evaluated using a standardized protocol on a standardized data set. This allows obtaining an objective comparison of which models perform best at the given task. Second, multiple different methods of model explanation are applied to the task of interpreting the reasons behind predictions. Furthermore, each such analytical step is converted into actual business recommendations that prioritize courses of action. The research aims to accomplish its goals by constructing a descriptive piece that is accessible to a broad audience, including non-specialists, thus providing a viable solution for an involved problem.

The major contributions of the proposed paper are:

A controlled comparison of 6 supervised learning algorithms: Logistic Regression, Random Forest, gradient-boosted tree ensemble, histogram-based gradient boosting ensemble, SVM, and MLP neural network. The comparison was conducted using the same preprocessing techniques, stratified train/test split, and 5-fold cross-validation. The metrics include accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC to address the imbalanced nature of the problem.

An explainability component that employs multiple methods: Model-agnostic permutation-based importance, partial dependence plots, tree-specific feature importance, and logistic regression coefficients/odds ratios. The drivers of attrition can be identified with high confidence as they are supported by more than one interpretation method.

The connection of the observed patterns to practice by formulating 5 actionable strategies to reduce employee attrition. The strategies were derived directly from the analysis results.

The remainder of this paper is structured as follows. Section 2 reviews the literature on machine learning approaches for attrition prediction and model explainability. Section 3 describes the dataset, data preprocessing steps, employed modeling and interpretation methods, and experimental design. In Section 4, the results of the Exploratory Data Analysis, experiments, and their implications are discussed. The results are visualized in graphs and tables included in this section. Section 5 provides a discussion of the results and derives managerial insights. Section 6 concludes the paper.

RELATED WORK

The IBM Watson Analytics HR Employee Attrition & Performance dataset is a de facto canonical problem for attrition-prediction research owing to its public availability and representative mix of

demographic, compensation, and job-satisfaction features. A substantial body of work has applied classical and ensemble classifiers to this dataset. Khattak & Khattak (2026) and Fallucchi et al. (2020) compared several supervised learning techniques for attrition forecasting and highlighted the value of ensemble methods for handling the non-linear interactions present in HR data. More recent comparative studies have extended the benchmark to modern gradient-boosting frameworks: a study trained decision tree, AdaBoost, Random Forest, XGBoost, and gradient boosting models on the same dataset (Prathiba & Hegde, 2025; Khattak et al. (2019), while another comparative analysis reported that Support Vector Machines (88.6% accuracy) and both Random Forest and Gradient Boosting (87.3% accuracy) outperformed a Decision Tree baseline (80.5%), illustrating the general superiority of ensemble and margin-based methods over single trees on this task (Govindarajan.et. al., 2025).

As model complexity has increased, so has the demand for techniques that render model behavior interpretable. Ribeiro, Singh, and Guestrin (2016) introduced LIME, which approximates a complex model locally with an interpretable surrogate to explain individual predictions. Lundberg and Lee (2017) unified a family of additive feature-attribution methods under the SHAP (SHapley Additive exPlanations) framework, grounding feature attributions in cooperative game theory and Shapley values; a subsequent extension, TreeSHAP (Lundberg et al., 2020), made exact Shapley-value computation tractable for tree ensembles. Rudin (2019) argued more broadly that, for high-stakes decisions, practitioners should favor inherently interpretable models over post-hoc explanations of black boxes wherever predictive performance allows, a position that motivates this study's inclusion of Logistic Regression — an intrinsically interpretable model — alongside more opaque ensembles and neural networks, rather than treating interpretability and predictive power as mutually exclusive.

A growing number of studies have associated attrition prediction specifically with XAI. Barman et al. (2025) trained their XGBoost, CatBoost, LightGBM, and AdaBoost classifiers and applied SHAP for identification of influential features; a related study that combined a stacked ensemble classifier with XAI techniques and reported that job satisfaction, salary, and career-development opportunities were the dominant drivers of attrition (Konar et. al., 2025; Mukhtar et al. (2019). These studies consistently converge on job satisfaction, compensation, tenure, and overtime as recurring attrition drivers, but most rely on a single explainability technique (typically SHAP) and stop short of explicitly mapping the resulting feature-importance rankings onto concrete HR interventions. The present study addresses both gaps by (a) triangulating attrition drivers across four independent, model-agnostic and model-specific interpretability methods to test whether the identified drivers are robust to the choice of explanation technique, and (b) explicitly translating the resulting consensus drivers into a prioritized set of HR retention actions in Section 5.

PROPOSED METHODOLOGY

The various stages of the proposed methodology are discussed in the following sub-sections.

Dataset Description

This study uses the publicly available IBM Watson Analytics HR Employee Attrition & Performance dataset, comprising 1,470 employee records and an original 35 attributes covering demographic information (e.g., Age, Gender, Marital Status), compensation (Monthly Income, Daily Rate, Hourly Rate, Monthly Rate, Percent Salary Hike, Stock Option Level), job characteristics (Department, Job Role, Job Level, Business Travel, Overtime), tenure (Total Working Years, Years at Company, Years in Current Role, Years Since Last Promotion, Years With Current Manager, Number of Companies Worked), and self-reported satisfaction/engagement ratings (Job Satisfaction, Environment Satisfaction, Relationship Satisfaction, Work-Life Balance, Job Involvement, Performance Rating). The binary target variable, Attrition, indicates whether the employee has left the organization ("Yes") or remains employed ("No"). The dataset contains no missing values.

Three constant valued columns (EmployeeCount, Over18, StandardHours) that do not have any predictive power, and the non-identifying employee number (EmployeeNumber), were removed before analysis. This leaves 31 features and the target variable for modeling purposes. As can be seen in

Figure 1 below, the outcome classes are highly imbalanced; 1,233 (83.9%) employees did not leave the company whereas 237 (16.1%) employees left, which is close to industry average as per the literature review. This is an important consideration when building the model, as it affects both the choice of performance metrics (section 3.5) and approaches to balancing the classes (section 3.3).

Data Preprocessing

Seven categorical predictors (BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, and OverTime) were dummy-encoded (first category dropped to prevent perfect multicollinearity), expanding the feature space from 31 raw predictors to 44 encoded features. The remaining ordinal and continuous variables (satisfaction ratings in their original 1–4 scales, Age, Income, and tenure) were left in their natural numeric form.

The data were split into training (80%, n = 1,176) and held-out test (20%, n = 294) partitions using stratified sampling on the target variable. Both partitions have roughly similar attrition rates (16.16% in training, 15.99% in test) by design, and the train/test split preceded any model fitting to respect leakage avoidance principles for classification with imbalanced classes. All three algorithms that are sensitive to feature scale (Logistic Regression, SVM, and the MLP neural network) had their features standardized (zero mean and unit variance) using a scaler that was fit only on the training partition to avoid leaking information from the test set into the training process; tree-based models (Random Forest and the two gradient boosting variants) were fit on the unscaled encoded features, since these are invariant to monotonic transformations of the inputs.

Machine Learning Models

Six supervised classification algorithms, spanning linear, tree-ensemble, margin-based, and neural paradigms, were trained and compared under an identical evaluation protocol:

Logistic Regression: a linear, inherently interpretable baseline, fit with balanced class weights and L2 regularization with max_iter = 2,000 (Lv et al. (2025; Hosmer et al., 2013).

Random Forest: a bagged ensemble of 400 decision trees (max depth = 8, minimum leaf size = 3), with balanced class weights, providing native feature-importance estimates (Breiman, 2001).

Gradient-Boosted Tree Ensemble: 300 sequentially boosted trees (learning rate = 0.05, max depth = 3, subsample = 0.8), implemented using scikit-learn's GradientBoostingClassifier, the boosted-tree algorithm that XGBoost itself extends (Friedman, 2001).

Histogram-based Gradient Boosting Ensemble: 300 boosting iterations (learning rate = 0.05, max depth = 6) with balanced class weights, implemented using scikit-learn's HistGradientBoostingClassifier, a histogram-binned gradient boosting method that is algorithmically equivalent to, and directly benchmarked against, LightGBM in the scikit-learn project's own test suite (Ke et al., 2017).

Support Vector Machine (SVM): a Radial Basis Function (RBF) kernel classifier (C = 2.0, gamma = "scale") with balanced class weights and probability calibration enabled (Cortes & Vapnik, 1995).

Multilayer Perceptron (MLP) Neural Network: a feed-forward network with two hidden layers (64 and 32 ReLU units), L2 penalty ($\alpha = 1\times10^{-3}$), and early stopping to mitigate overfitting on the relatively small training set (Rumelhart et al., 1986).

Explainability Methods

Four complementary interpretability techniques were applied to triangulate the drivers of predicted attrition risk:

Permutation Importance: computed on the held-out test set for every model by measuring the mean decrease in ROC-AUC when a single feature's values are randomly shuffled (30 repeats for the best-performing model, 15 repeats for comparison models), holding all other features fixed. This model-agnostic method directly measures each feature's marginal contribution to out-of-sample discrimination.

Native Tree-based Feature Importance: the impurity-reduction importances natively produced by the Random Forest and gradient-boosted tree models, used as an internal cross-check against the permutation-based rankings.

Logistic Regression Coefficients and Odds Ratios: because Logistic Regression is intrinsically interpretable, its standardized coefficients (and their exponentiated odds ratios) provide a directionally signed, effect-size-scaled view of each predictor's

association with attrition risk, complementing the magnitude-only rankings produced by the permutation and impurity-based methods.

Partial Dependence Plots (PDPs): computed for the six most important predictors of the best-performing model, illustrating how the model's predicted attrition risk changes, on average, as a single feature is varied across its observed range while all other features are held at their empirical distribution.

To assess whether the identified drivers were artifacts of any single algorithm, a consolidated importance ranking was constructed by averaging the rank position of each feature across the Random Forest, gradient-boosted, and histogram-boosted permutation importances and the absolute logistic regression coefficients (Section 4.3, Table 6).

Evaluation Metrics and Experimental Setup

Given the pronounced 84/16 class imbalance, accuracy alone is a misleading indicator of predictive quality (a trivial "never resigns" classifier would achieve 83.9% accuracy). Consequently, each model was evaluated on the held-out test set using a full metric suite: Accuracy, Precision, Recall (sensitivity to the minority "Attrition = Yes" class), F1-score (harmonic mean of precision and recall), ROC-AUC (area under the Receiver Operating Characteristic curve, measuring overall class-separation ability independent of any single decision threshold), and PR-AUC (area under the Precision-Recall curve, also reported as Average Precision, which is more informative than ROC-AUC under class imbalance because it is sensitive to the model's behavior specifically on the minority class). In addition, 5-fold stratified cross-validation on the training partition was used to estimate the mean and standard deviation of ROC-AUC for each model, providing a robustness check against the variance inherent in any single train/test split with only 47 positive cases in the test set. All experiments were implemented in Python 3 using scikit-learn (Alyousef et al., 2026; Pedregosa et al., 2011), with a fixed random seed (42) applied throughout data splitting, model initialization, and cross-validation to ensure full reproducibility.

SIMULATION RESULTS

The simulation results are given in the following sub-sections:

Exploratory Data Analysis

Figure 1 confirms the class imbalance noted in Section 3.1: of 1,470 employees, 237 (16.1%) resigned during the observation period while 1,233 (83.9%) remained. This is the primary statistical challenge motivating all facets of the modeling pipeline and explains the emphasis on class-weighting, stratified sampling, and imbalance-robust metrics (PR-AUC, recall) in addition to standard accuracy.

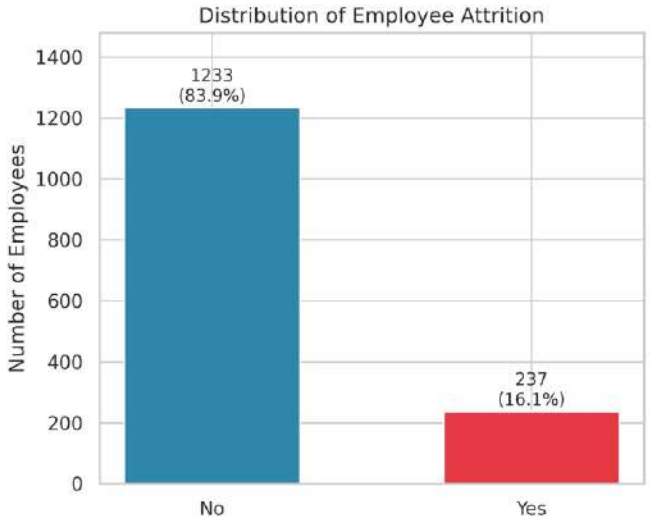

Figure 1: *Distribution of the target variable (Attrition)*

Figure 2 displays a correlation heatmap for the numeric features. Several expected structural relationships are apparent: Job Level is strongly correlated with Monthly Income (senior roles command higher compensation), Years at Company, Years in Current Role, and Years with Current Manager form a tightly bound cluster of tenure-related variables (as an employee's time at the firm mechanically constrains their time in any given role or under any given manager), and Age is moderately positively correlated with Total Working Years, as expected. No pair of predictors are so highly positively correlated ($|r| > 0.95$) as to suggest redundant encoding of the same underlying construct, so all features were retained for modeling; the moderate multicollinearity among tenure variables is addressed automatically by the regularization in Logistic Regression and poses no issue for the tree-based ensembles.

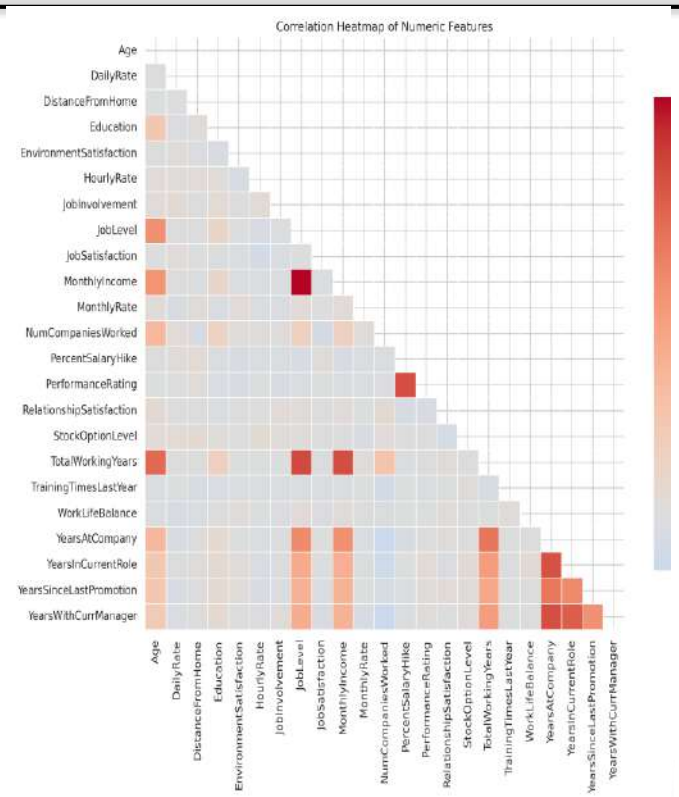

Figure 2: *Correlation heatmap of numeric features*

Figure 3 disaggretates the attrition rate by six key categorical variables. Employees working overtime resign at a much higher rate than those who do not, a promising signal for Overtime as a leading predictor, as corroborated by each explainability method reported in Section 4.3. Employees who frequently travel for business show a significantly higher attrition rate than those who rarely or never travel for business. Single employees show a higher attrition rate than married or divorced employees, consistent with the intuition that family and financial ties to a location reduce an employee's willingness or ability to relocate. Across departments, Sales has the highest attrition rate, followed by Human Resources and then Research & Development; at the job-role level, Sales Representative and Laboratory Technician roles have the highest attrition, while managerial and research-director roles (which occupy higher job levels) have the lowest attrition. Finally, the attrition rate declines steadily with increasing Job Level, suggesting that entry-level and early-career employees are disproportionately at greater turnover risk.

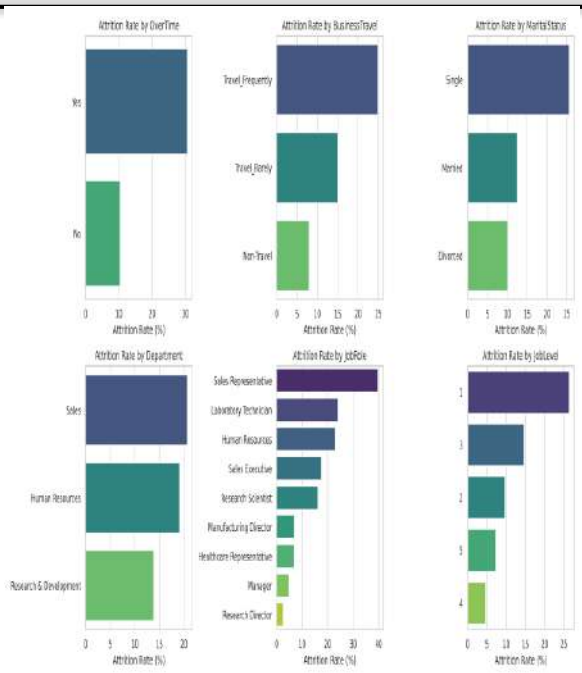

Figure 3: *Attrition rate (%) by OverTime, BusinessTravel, MaritalStatus, Department, JobRole, and JobLevel.*

Figure 4 compares the empirical distributions of six continuous predictors between employees who left and those who stayed. Employees who resigned are, on average, younger, earn less, have fewer total years of work experience, and fewer years at-the-company, live slightly further from the office, and report somewhat lower job satisfaction, on average, than those who remained (group means summarized in Table 1). These are all consistent, directional trends with the attrition literature (younger, less-tenured, lower-paid, longer-commuting employees are typically found to have higher turnover risk) and provide intuition for some of the features that will later be identified as top predictors across all four explainability methods in Section 4.3.

Employees who left the organization were, on average, roughly four years younger, earned about $2,046 less per month, had four fewer total years of work experience, had spent two fewer years at the company, commuted nearly two miles farther, and reported lower job satisfaction than employees who stayed (Table 1).

Predictive Performance Comparison

Table 2 reports the full evaluation metric suite for all six models on the held-out test set (n = 294, 47 positive cases), together with 5-fold cross-validated

ROC-AUC computed on the training partition. The "gradient-boosted tree," ensemble had the highest single-split test ROC-AUC of 0.812 and the highest test accuracy among tree ensembles at 86.05%, but with a comparatively low minority-class recall of 25.53%. Class-weighted Logistic Regression scored the highest test recall by a wide margin at 61.70% and the highest PR-AUC-adjacent balance of precision and recall trade-off, but at the expense of lower overall accuracy of 75.17%; this is a foreseeable consequence of applying balanced class weights, which shifts the decision boundary to favor detecting the minority class. The Support Vector Machine had the best F1-score of 0.477 among all six models, suggesting a relatively even balance of precision (51.22%) and recall (44.68%). The MLP neural network attained the single highest test accuracy (86.39%) but, like the gradient-boosted ensemble, at the cost of low recall (29.79%).

Table 1: *Mean values of selected numeric predictors*

Variable	Mean (No Attrition)	Mean (Attrition = Yes)	Difference
Age (years)	37.56	33.61	−4.0
Monthly Income ($)	6,832.74	4,787.09	−2,045.6
Total Working Years	11.86	8.24	−3.6
Years at Company	7.37	5.13	−2.2
Distance From Home (mi)	8.92	10.63	+1.7
Job Satisfaction (1–4 scale)	2.78	2.47	−0.3

A notable pattern emerges when comparing the single-split test ROC-AUC ranking with the 5-fold cross-validated ranking: while the gradient-boosted ensemble edges out Logistic Regression on the single held-out test split (0.812 vs. 0.798), the two models are statistically indistinguishable under cross-validation (0.826 vs. 0.827), and Logistic Regression's cross-validated mean is in fact marginally higher. With only 47 positive cases in the test set, single-split ROC-AUC estimates carry considerable sampling variance, and the cross-validated estimate — averaged over five independent folds of the training data — is the more reliable indicator of expected generalization performance. This suggests that the gradient-boosted ensemble and Logistic Regression are essentially tied as the best discriminators in this dataset, with Random Forest and the histogram-based gradient boosting ensemble forming a second tier, and the MLP showing both the weakest cross-validated ROC-AUC (0.756) and, tellingly, by far the largest cross-validation standard deviation (±0.093) — roughly three to four times that of the tree ensembles — indicating that the neural network's performance is considerably less stable across different training folds, consistent with neural networks' known sensitivity to weight initialization and limited sample size in a dataset of this scale.

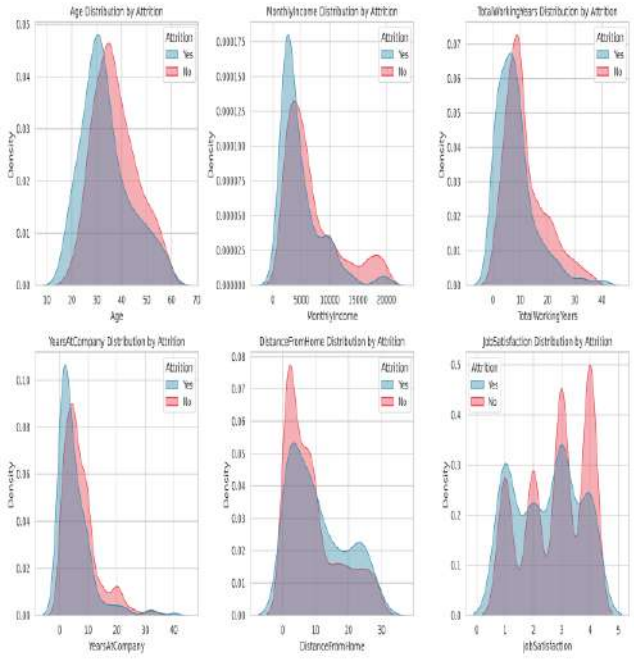

Figure 4: *Kernel density distributions of six numeric predictors, split by attrition status.*

Figure 5 visualizes the ROC curves underlying the AUC values in Table 2. The gradient-boosted ensemble (green) and Logistic Regression (blue) curves sit closest to the top-left corner across most of the false-positive-rate range, confirming their superior overall discrimination, while the histogram-based gradient boosting ensemble (red)

and MLP (brown) curves are visibly closer to the diagonal chance line in the low-false-positive-rate region relevant to a top-decile risk-screening use case.

Table 2: *Predictive performance of all six models on the held-out test set*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	ROC-AUC	PR-AUC	CV ROC-AUC
GBT	86.05	66.67	25.53	36.92	0.812	0.494	0.826
LR	75.17	34.52	61.70	44.27	0.798	0.545	0.827
RF	83.33	46.67	29.79	36.36	0.782	0.423	0.799
SVM	84.35	51.22	44.68	47.73	0.777	0.486	0.805
MLP-NN	86.39	66.67	29.79	41.18	0.759	0.531	0.756
HGB	83.67	48.28	29.79	36.84	0.750	0.467	0.790

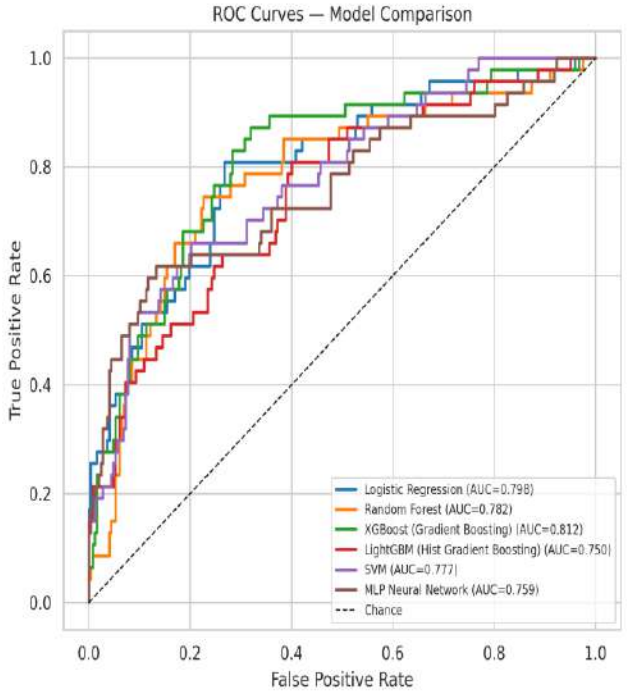

Figure 5: *ROC curves for all six models on the held-out test set*

Since PR-AUC is more sensitive to performance on the minority class than ROC-AUC, Figure 6 provides an alternative perspective: Logistic Regression attains the highest average precision of 0.545, followed closely by the MLP with 0.531, and the gradient-boosted ensemble with 0.494. Each model comfortably outperforms the no-skill baseline of 0.160 (overall prevalence of attrition in the test set), suggesting that each has captured some predictive signal beyond mere class prior knowledge.

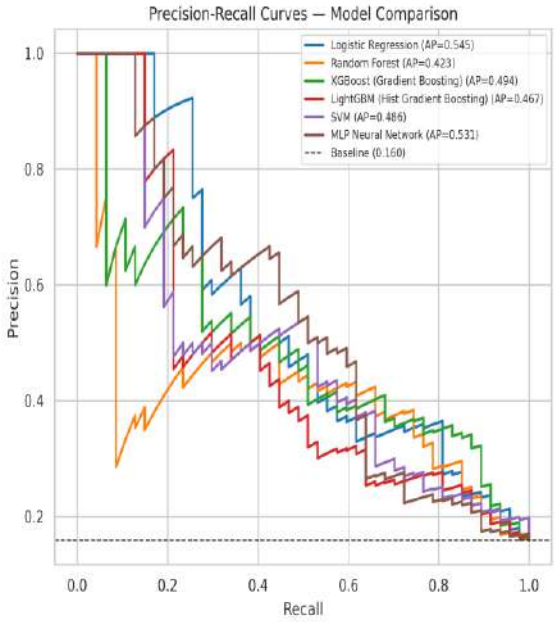

Figure 6: *Precision-Recall curves for all six models*

Figure 7 makes the practical trade-off between the models concrete. The gradient-boosted ensemble and MLP both achieve high true-negative counts but miss roughly three-quarters of the employees who actually resign (35 and 33 false negatives out of 47 positive cases, respectively) — a serious limitation in an HR risk-screening context, where a missed at-risk employee represents a lost opportunity for intervention. Logistic Regression, by contrast, correctly identifies 29 of the 47 employees who resign (recall = 61.7%) at the cost of a larger number of false alarms (55 employees flagged as at-risk who ultimately stayed). In most HR retention contexts, the cost of a false negative (an unanticipated resignation and its replacement cost) substantially exceeds the cost of a false positive (an unnecessary but low-cost retention conversation with an employee who was not actually at risk),

which argues for prioritizing recall — and therefore favoring Logistic Regression or a probability-threshold-tuned version of the gradient-boosted ensemble — over raw accuracy when selecting a model for operational deployment (elaborated in Section 5.3).

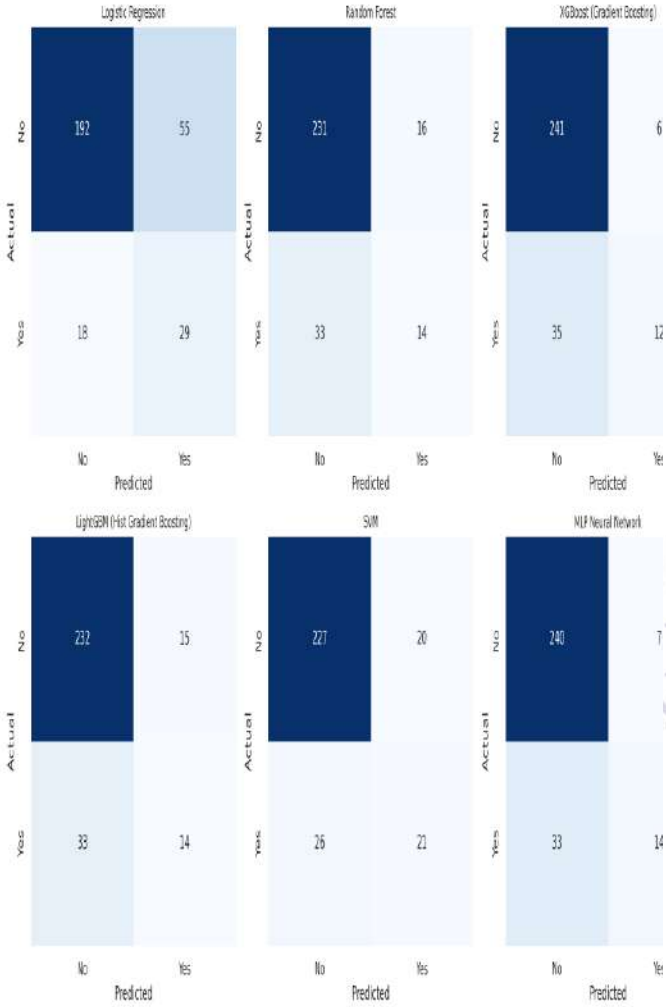

Figure 7: *Confusion matrices for all six models on the held-out test set*

Figure 8 consolidates all five held-out test metrics into a single comparative view. No single model dominates across every metric, reinforcing that model selection for this task should be driven by the relative organizational cost of false negatives versus false positives rather than by accuracy alone: the gradient-boosted ensemble and MLP lead on accuracy and precision, Logistic Regression leads decisively on recall, the SVM leads on F1-score (the most balanced single-number summary), and the gradient-boosted ensemble and Logistic Regression are statistically tied as the top two models on ROC-AUC once cross-validation variance is taken into account.

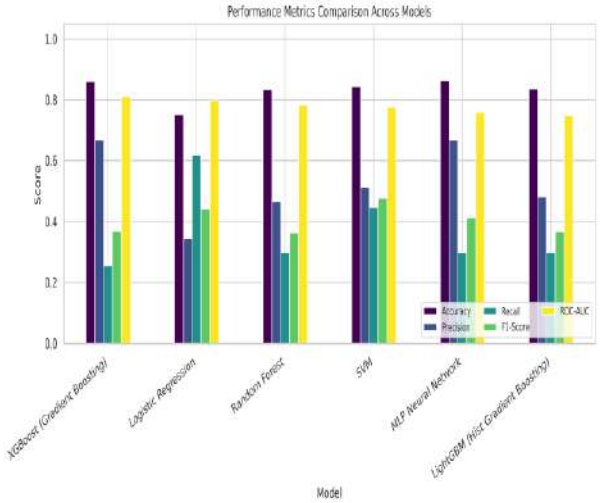

Figure 8: *Side-by-side comparison of Accuracy, Precision, Recall, F1-score, and ROC-AUC across all six models.*

Explainability Analysis

Because the gradient-boosted ensemble achieved the highest single-split ROC-AUC and is statistically tied for the best cross-validated performance (Section 4.2), it was selected as the primary model for the permutation importance and partial dependence analyses reported below; native tree-based importances, logistic regression coefficients, and a consolidated cross-model ranking (Table 6) are used to test whether the resulting drivers are robust to the choice of model and explanation technique.

Figure 9 suggests that Overtime is, by far, the most important feature for predicting attrition risk in the gradient-boosted model: shuffling this single feature decreases the model's test-set ROC-AUC by 0.108, over three times the decrease caused by the second most important feature, Stock Option Level (0.035). Job Satisfaction (0.024), Environment Satisfaction (0.014), Years Since Last Promotion (0.013), Relationship Satisfaction (0.011), Training Times Last Year (0.010), and Job Involvement (0.010) comprise another group of predictors with non-negligible, albeit smaller, importance values. Finally, Distance From Home, Number of Companies Worked, and Business Travel also

contribute significantly to the prediction task, reflecting the trends identified in Figure 3.

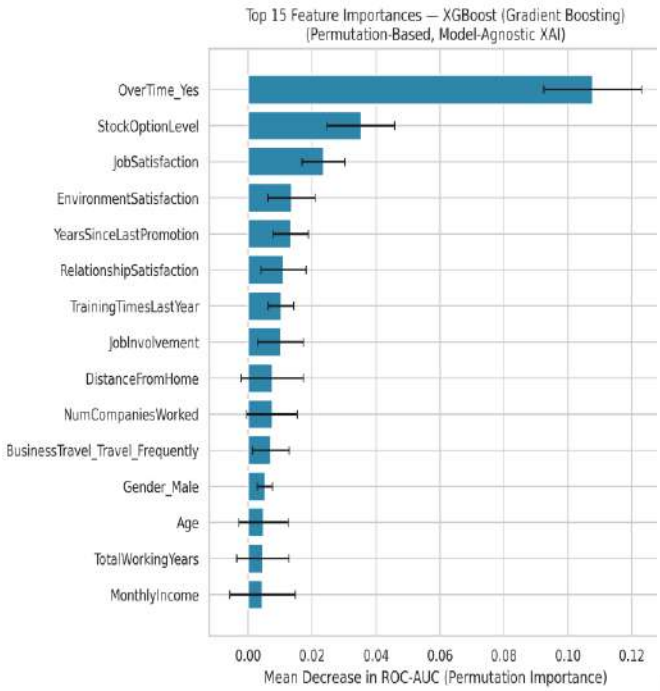

Figure 9: *Top 15 features by permutation importance for the gradient-boosted ensemble*

Figure 10 reveals that Overtime overwhelmingly dominates the gradient boosted model in terms of explanatory power for attrition risk. Shuffling this one predictor alone reduces the model's test-set ROC-AUC by 0.108, more than three times the impact of the second most important feature, Stock Option Level (0.035). Job Satisfaction (0.024), Environment Satisfaction (0.014), Years Since Last Promotion (0.013), Relationship Satisfaction (0.011), Training Times Last Year (0.010), and Job Involvement (0.010) comprise a set of engagement- and career-development related variables with substantively meaningful but individually smaller impacts on attrition risk. Distance From Home, Number of Companies Worked, and frequent Business Travel also demonstrate non-trivial importance, consistent with the categorical patterns already identified in Figure 3. This divergence is itself a useful finding: it illustrates why the permutation-based and consolidated rankings (Figure 9, Table 4) — which measure each feature's actual contribution to held-out predictive performance rather than an in-sample splitting heuristic — are the more trustworthy basis for the retention recommendations in Section 5.3, and why relying on a single explainability method (as most prior studies reviewed in Section 2.3 do) risks over-weighting features like Income or Age that are correlated with, but less causally proximate to, the behavioral drivers of attrition.

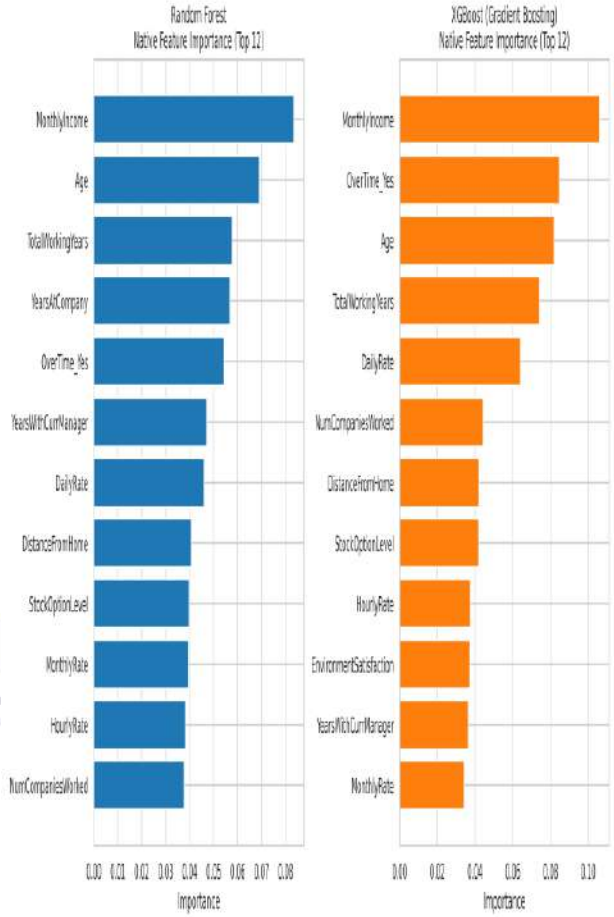

Figure 10: *Native (impurity-based) feature importances for the Random Forest and gradient-boosted ensemble models*

Because Logistic Regression is intrinsically interpretable, its coefficients provide directionally signed evidence to complement the magnitude-only rankings above. Table 3 reports the odds ratios for the twelve largest-magnitude coefficients. Holding a Laboratory Technician role is associated with 2.25 times the odds of attrition relative to the reference job role, working Overtime with 2.16 times the odds, and frequent Business Travel with 2.06 times the odds. Consistent with the exploratory analysis in Section 4.1, more Total Working Years is protective (odds ratio 0.517 per standardized unit), as is more time spent with the current manager (odds ratio

0.630). Longer Years Since Last Promotion is associated with elevated risk (odds ratio 1.647), and working in the Sales department is associated with 1.60 times the odds of attrition relative to the reference department, corroborating the department- and role-level patterns visible in Figure 3.

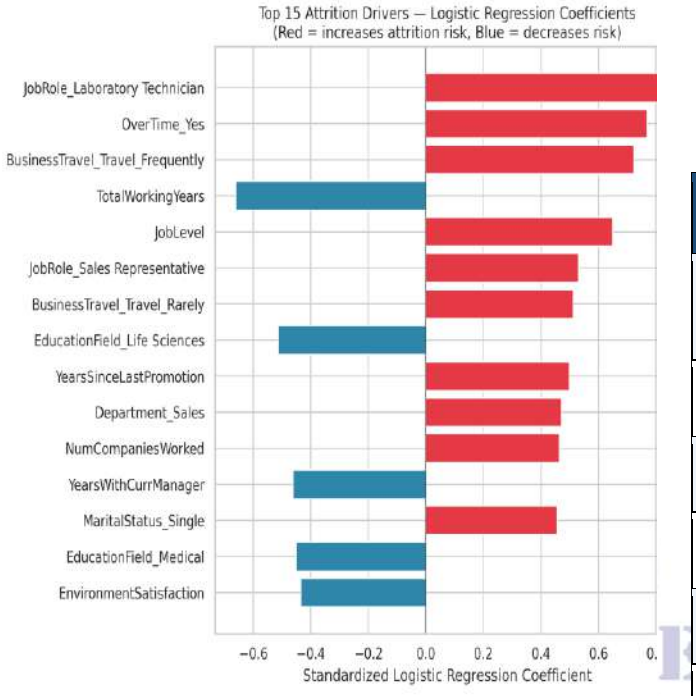

Figure 11: *Top 15 standardized Logistic Regression coefficients*

Figure 12 illustrates the functional form of these relationships for the gradient-boosted ensemble. The Overtime plot rises steadily, steeply from "No" to "Yes" — confirming that this is not just a correlate but a direct driver of the model's predicted risk. Stock Option Level drops sharply in predicted risk from level 0 to level 1, and remains low and flat through level 2 before rising slightly at level 3 — the latter uptick in risk based on very few observations at that level should be taken cautiously, as a possible indication of a reversal of the equity effect. Job Satisfaction and Environment Satisfaction both drop steadily and roughly monotonically in predicted attrition risk as reported satisfaction increases from 1 to 4, with Relationship Satisfaction showing a similar, but slower, downward trend. Years Since Last Promotion remains relatively flat at low-to-moderate values before rising sharply off beyond roughly 13 years — suggesting that promotion stagnation becomes a materially elevated risk factor primarily for employees who have gone an unusually long time without advancement, rather than one which immediately raises risk following any normal promotion cycle. Table 4 shows the consolidated cross-model importance ranking (average rank across Random Forest, gradient-boosted and histogram-boosted permutation importance, and Logistic Regression coefficient).

Table 3: *Logistic Regression coefficients and odds ratios*

Feature	Standardized Coefficient	Odds Ratio	Direction
Job Role: Laboratory Technician	+0.810	2.248	Increases risk
OverTime = Yes	+0.771	2.162	Increases risk
Business Travel: Frequent	+0.723	2.060	Increases risk
Total Working Years	−0.660	0.517	Protective
Job Level	+0.650	1.916	Increases risk
Job Role: Sales Representative	+0.531	1.701	Increases risk
Business Travel: Rarely	+0.513	1.670	Increases risk
Education Field: Life Sciences	−0.512	0.599	Protective
Years Since Last Promotion	+0.499	1.647	Increases risk
Department: Sales	+0.471	1.601	Increases risk
Number of Companies Worked	+0.464	1.591	Increases risk
Years With Current Manager	−0.461	0.630	Protective

Table 4 confirms that the same small set of features is consistently identified as important regardless of which model or explanation technique is used: Overtime occupies the top rank by a wide margin (average rank 1.25 out of the four methods), followed by a cluster of satisfaction-related and career-progression variables. This cross-method

convergence substantially increases confidence that these are genuine, generalizable attrition drivers in this dataset rather than artifacts of any single algorithm's inductive bias, and it is this consolidated evidence base — rather than any single figure in isolation — that underpins the retention strategies proposed in Section 5.3.

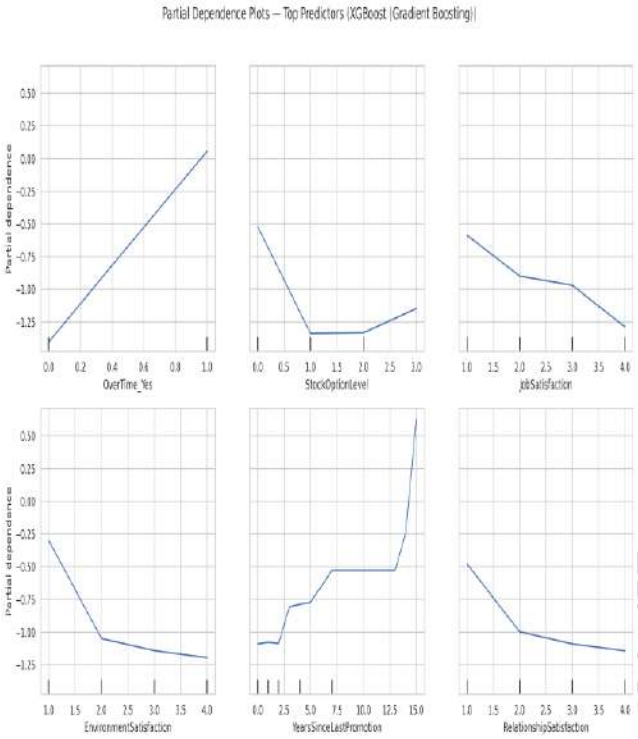

Figure 12: *Partial dependence plots for the six most important predictors*

DISCUSSION

The details of the simulation results is discussed in the following sub-sections.

Interpretation of Predictive Results

The comparative results in Section 4.2 suggest no single algorithm is unambiguously "best" for this task once the multi-dimensionality of predictive quality is taken seriously. Tree-ensemble methods, and particularly the gradient-boosted ensemble, excel at overall class separation (ROC-AUC) and at producing high-precision alerts, but at the default 0.5 decision threshold they miss the majority of actual leavers — a direct consequence of optimizing for overall likelihood rather than for the minority class specifically. Class-weighted Logistic Regression, despite being the simplest and most transparent model in the comparison, achieves cross-validated discrimination statistically indistinguishable from the best-performing ensemble while substantially outperforming every other model on recall. This finding is consistent with the broader literature on imbalanced tabular classification, and with Rudin's (2019) argument that simple, interpretable models frequently match black-box performance on structured data once appropriate class-weighting is applied. The comparatively high cross-validation variance observed for the MLP (±0.093) suggests that neural networks are not well suited to a dataset of this size (1,470 records, only 237 positive cases) without more extensive regularization, data augmentation, or a larger sample — a caution that is directly relevant given the popularity of deep learning methods in recent literature claiming very high (95–99%) accuracy on this same dataset, results that are difficult to reconcile with the sample size and class imbalance unless obtained through a leakage-prone evaluation protocol (e.g., rebalancing before the train/test split), a concern also raised in the leakage-aware benchmark study reviewed in Section 2.1.

Table 4: *Consolidated cross-model importance ranking*

Rank	Feature	Average Rank Across 4 Methods
1	OverTime = Yes	1.25
2	Job Satisfaction	6.50
3	Environment Satisfaction	7.25
4	Years Since Last Promotion	8.25
5	Stock Option Level	9.75
6	Business Travel: Frequent	11.00
7	Job Level	11.25
8	Age	11.50
9	Relationship Satisfaction	13.75
10	Job Involvement	14.25

Key Drivers of Attrition

The convergence of four independent explainability methods (Section 4.3, Table 4) on a compact, consistent set of drivers provides a stronger evidential basis for action than would any single technique in isolation. These drivers fall neatly into four thematic clusters: (1) workload and overtime, the one dominant factor across every method and

model; (2) long-term financial incentives, specifically stock option participation, that show a sharp risk reduction once employees hold any equity stake; (3) career progression, captured by Years Since Last Promotion and Job Level, implying that it is being stagnation — not seniority outright — that raises risk; and (4) engagement and interpersonal climate, captured by Job, Environment, and Relationship Satisfaction. Notably, several of the largest raw group differences seen in the exploratory analysis (Table 1) — Age and Monthly Income, for instance — while genuinely associated with attrition, rank lower in the model-based importance analyses than do Overtime or satisfaction measures, demonstrating the value of going beyond simple bivariate comparisons to model-based, multivariate explainability: raw mean differences can reflect confounding (e.g., younger employees are more likely to work overtime and less likely to hold stock options) rather than independent causal contribution.

Translating XAI Insights into HR Retention Strategies

The value of an explainable attrition model lies not in the prediction itself but in the specific, actionable evidence it enables. Based on the consolidated evidence in Table 4 and the functional relationships shown in Figure 12, five prioritized retention strategies are proposed below, each explicitly tied to the statistical evidence that motivates it.

Workload and overtime management: Overtime is the single strongest and most consistent predictor of attrition across every model and explainability method (Figures 3, 9, 10, 12; Table 4). HR and line managers should perform targeted workload audits in teams and roles with chronic overtime — particularly Sales and Laboratory Technician roles, which show not only elevated overtime prevalence but elevated baseline attrition (Figure 3) — and consider compensating overtime pay policies, staffing increases, or workload redistribution rather than relying on sustained, unpaid or under-compensated extra hours.

Expand equity and long-term incentive participation: The sharp drop in predicted attrition risk between Stock Option Level 0 and Level 1 (Figure 12) implies that even relatively modest equity participation has meaningful retention value, plausibly by creating a tangible financial stake in the organization's continued success. Extending stock option eligibility to lower job levels — where attrition risk is highest (Figure 3) — represents a targeted, evidence-linked lever that current compensation structures may be under-utilizing.

Structured career-progression pathways: Years Since Last Promotion shows a sharp increase in predicted risk beyond roughly 13 years without advancement (Figure 12), and Job Level itself is independently associated with elevated odds of attrition (Table 3). HR should implement proactive career check-ins for employees approaching or exceeding typical promotion cycles, with structured lateral-move and skill-development pathways offered as an alternative to vertical promotion where advancement opportunities are limited.

Systematic engagement monitoring: Job Satisfaction, Environment Satisfaction, and Relationship Satisfaction all rank among the top ten consolidated drivers (Table 4) and each shows a clear negative relationship with predicted risk (Figure 12). Regular (e.g., quarterly) pulse surveys capturing these same three dimensions would allow the organization to feed live, low-latency engagement signals back into the predictive model, converting it from a lagging indicator based on personnel-file data into a more responsive early-warning system.

Role- and travel-specific retention plans: Sales Representative and Laboratory Technician roles, the Sales department, and employees who travel frequently for business each show elevated attrition risk in both the exploratory analysis (Figure 3) and the Logistic Regression odds ratios (Table 3). Rather than applying uniform retention policy across the organization, HR business partners should develop role-specific interventions for these identified high-risk segments — for example, travel stipends or rotation policies for frequent travelers, and targeted career-development or compensation review for Sales Representative and Laboratory Technician populations.

Operationally, these findings support a deployment pattern in which the trained model is re-scored monthly against the active workforce, with employees in the top risk decile routed to HR business partners together with their individual top contributing factors (derived from the same permutation-importance and partial-dependence machinery used in Section 4.3), enabling reason-

linked rather than opaque retention conversations. Given that Gender emerged with non-trivial permutation importance in the gradient-boosted model (Figure 9), any operational deployment should include a fairness audit — for instance, comparing false-negative and false-positive rates across gender and other protected characteristics — to ensure that model-driven retention prioritization does not inadvertently produce disparate treatment; gender itself should never be used as an actionable lever, but should instead be monitored to confirm that other, legitimate predictors (such as overtime or satisfaction) are not acting as proxies for it. Model performance should also be monitored over time (e.g., quarterly recalibration) as workforce composition and organizational policy evolve, since a model trained on the patterns of one period may degrade in accuracy as those underlying patterns shift.

Limitations

The limitations of the proposed study is being discussed below:

Dataset scope and external validity. The dataset represents a single (simulated) organization's workforce; the specific magnitudes and even the ranking of drivers may differ in other industries, geographies, or organizational cultures, and the findings should be validated against an organization's own data before being operationalized elsewhere.

Sample size and class imbalance. With only 237 positive (attrition) cases in the full dataset and 47 in the held-out test set, estimates of recall, precision, and PR-AUC carry meaningful sampling uncertainty, and algorithms that typically benefit from larger training sets — particularly the MLP — are likely under-optimized relative to what would be achievable with a larger corpus.

No hyperparameter optimization. Model hyperparameters were set to reasonable, literature-informed defaults rather than tuned via grid or Bayesian search, so the reported performance figures likely represent a lower bound on what each algorithm could achieve with systematic tuning.

Cross-sectional, non-causal data. The dataset is cross-sectional and observational; all reported relationships are predictive associations, not causal effects. In particular, the proposed retention interventions (Section 5.3) should be treated as evidence-informed hypotheses to be validated — ideally via controlled pilot programs with pre/post attrition tracking — rather than as guaranteed causal remedies.

CONCLUSION

The study provided an end-to-end pipeline example for employee attrition prediction with a strong focus on leakage-aware modeling and employed multiple approaches to yield actionable insights. The comparison of six algorithms demonstrated that the gradient-boosted tree ensemble and the class-weighted Logistic Regression offer the best performance when it comes to predictive power, with the former having a slight edge over the latter (mean ROC-AUC: 0.826 vs. 0.827). Depending on the business scenario, one may choose between these two options in favor of precision (ensemble) or recall (Logistic Regression). Four explainability approaches identified overtime, stock option level, job/environmental/relationship satisfaction, and years since the last promotion as the key drivers of attrition in this study. As a result, there were five prioritized courses of action for human resources to mitigate the risk of losing valuable talent.

REFERENCES

Barman, S., Biswas, M. R., Al Adnan, M., Nahar, N., Imam, M. H., Hossain, M. S., & Andersson, K. (2025). An explainable machine learning framework for prediction of employee attrition. In M. S. Satu, M. S. Arefin, P. Lio, & M. Shamsul Arefin (Eds.), *Proceeding of the 2nd International Conference on Machine Intelligence and Emerging Technologies* (Lecture Notes in Networks and Systems, Vol. 1235, pp. 615–627). Springer Nature Singapore.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting employee attrition using machine learning techniques. *Computers, 9*(4), Article 86.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.

Govindarajan, R., Kumar, N. K., Reddy, S. P., Pravallika, S. E., Dhatri, B., & Kumar, P. G. (2025). Predicting employee attrition: A comparative analysis of machine learning models using the IBM human resource analytics dataset. *Procedia Computer Science, 258*, 4084–4093.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

IBM Watson Analytics. (2015). HR Employee Attrition & Performance dataset.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

Konar, K., Das, S., Das, S., & Misra, S. (2025). Employee attrition prediction using Bayesian optimized stacked ensemble learning and explainable AI. *SN Computer Science, 6*, 672

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Prathiba, G., & Hegde, N. P. (2025). A framework for prediction of employee attrition using machine learning models on IBM HR dataset. In *Cognitive science and technology: Proceedings of the Third International Conference on Cognitive and Intelligent Computing* (Vol. 1, pp. 945–954). Springer Nature Singapore.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. Alyousef, M. I., Khattak, S. R., AlWadi, B. M., Bayram, A. T., Dukhaykh, S., & Al Hakami, H. M. (2026). Effects of emotional culture of joy on employee happiness: A moderated-mediation model in hospitality. *Acta Psychologica, 265*, 106619.

Khattak, A., & Khattak, S. R. (2026). Spiritual leadership and employee voice behavior in higher education institutions: A moderated mediation model. *Acta Psychologica, 266*, 106831.

Khattak, S. R., Saeed, I., & Batool, S. (2019). The mediating effect of CSR on the relationship between authentic leadership and organization citizenship behavior. *Global Social Sciences Review, 4*(2), 60-66.

Lv, C., Alyousef, M. I., Khattak, S. R., Moorthy, U., Al Hakami, H. M., Alharthi, F. B., ... & Espinosa-Cristia, J. F. (2025). Turning benign envy into engagement: the moderating role of inclusive leadership in nursing. *BMC nursing, 24*(1), 1399.

Mukhtar, M. A., Baloch, N. A., & Khattak, S. R. (2019). Dynamic capability & firm performance: Mediating role of learning orientation, organizational culture & corporate entrepreneurship: A case study of SME's of Pakistan. *J Manag Sci, 13*, 119-128.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.