International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(2): 862-873 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i02.03512

Featuring Machine Learning Models to Evaluate Employee Attrition: A Comparative Analysis of Workforce Stability-Relating Factors

Mustafizul Haque[1], Tejasvini Alok Paralkar[2], Sudhir Rajguru[3], Adheer A Goyal[4], Tanaya Patil[5], Kamal Upreti[6]

[1]Centre for Online Learning, Dr. D. Y. Patil Vidyapeeth, Pune (Deemed to be University), India, [2]Symbiosis Centre for Management Studies, Symbiosis International (Deemed University) Nagpur Campus, India, [3]Department of Management, IIMT, Greater Noida, India, [4]School of Commerce & Management, GH Raisoni University, Saikheda, India, [5]Department of Management, Sanjivani University Maharashtra, India, [6]Department of Computer Science, CHRIST (Deemed to be University), Delhi NCR, Ghaziabad, India. *Corresponding Author's Email: kamalupreti1989@gmail.com

Abstract

Employee attrition is a problem for most organizations as it affects morale, productivity, and business continuity. In addressing this, the study made use of machine learning techniques such as Clear AI, Random Forest, and logistic regression in designing a prediction model to predict who is the next to leave within an organization. The HR data relating to demographics, performance metrics, job roles, and conditions of work was sourced from publicly available website Kaggle.com for the study. Data preprocessing included scaling, outlier detection, and balancing the dataset using SMOTE. Multiple machine learning models were trained and evaluated by checking on accuracy, F1-score, and the ROC-AUC curve. The best model that was tested was Random Forest, which gave an accuracy of 85.71%. Additional insights from feature importance highlighted the significant effect of overtime, marital status, and stock options on attrition. Among the remaining key drivers are workload, work-life balance, and financial incentives. These findings suggest the need for focused HR strategies, such as reduction of overtime, mentorship programs, and career development opportunities, to reduce attrition rates and improve employee satisfaction. This study provides a robust methodology in predicting attrition and delivers actionable insights into designing interventions that improve workforce stability and organizational efficiency.

Keywords: Employee Attrition, Features Importance, Human Resources, Machine Learning Models, Organizations.

Introduction

Employee attrition or turnover is a serious issue that organizations face globally, affecting their productivity, workforce stability, and the success of the organization as a whole (1, 2). Attrition refers to the rate at which employees leave an organization, either voluntarily or involuntarily, and this is often a significant challenge for human resources teams to retain valuable talent (3-5). High turnover rates result in business processes disruptions, high hiring and training costs, and the loss of institutional knowledge. Such disruptions may affect company morale and performance (6-8). Various factors are involved in employee attrition, including job satisfaction (9), career growth opportunities (10), compensation (11), work-life balance (12), and interpersonal relationships within the workplace (13). External elements like economic uncertainty and lately, the COVID-19 pandemic have brought new aspects to the workplace, like working remotely and changing employees' expectations (14). Despite a great number of researches in employee attrition, the significant gap remains unattended: it is about understanding how the impacts of both internal and external factors might interplay together in a specific context-specific scenario, for example, specific industries or regions (15, 16). Current research on attrition tends to be on predictive models; however, the current challenge remains enhancing the transparency and interpretability of such models, which would help HR professionals translate these findings into actionable decision-making tools. This problem is further worsened by increasing diversity and complexity in the needs of employees, along with evolving workplace dynamics resulting from technological advance-

(Received 13th December 2024; Accepted 15th April 2025; Published 30th April 2025)

-ments and global health crises.

High levels of attrition do not only generate high recruitment and training expenses but also result in loss of knowledge and decreased organizational morale. It is with increasing necessity that the current study draws motivation to derive data-driven tactics that enable organizations to identify problem employees early on and develop precision retention interventions. Through the utilization of interpretable machine learning models, this work seeks to enable HR professionals to make informed decisions with actionable findings to tackle one of the most critical workforce management issues.

The objective is to develop a model to predict high-risk employee turnover in an Iranian pharmaceutical company (17). Human resource manager interviews, sophisticated data mining tools, and data from the human resources department are all employed in this combined research strategy. With 89% accuracy, the algorithm forecasts high-risk attrition. The study also emphasizes how COVID-19 and Employee attrition is impacted by remote work. Using 30 attributes from IBM HR Analytics Employee Attrition and Performance data, this study proposes a predictive model for employee attrition predictions. Eight prediction models were developed and evaluated by researchers: ensemble model, XGBoost, SVM, Random Forest, Logistic Regression, and Artificial Neural Network model (18). According to the survey, the main causes of employee attrition were relationship satisfaction, environmental contentment, and overtime work. This emphasizes how crucial talent management is to lowering employee turnover. To address the issue, the use of explainable AI (XAI) to predict employee turnover and offer data-driven remedies was also explored (19). It looks at how attrition affects output, morale, and financial security as well as how AI can forecast staff departures by using past data and employee behavior. In order to help HR professionals, create focused recruiting and retention strategies, the article also presents explainable AI approaches, which increase the transparency and interpretability of AI models. Machine learning techniques are used to forecast IBM employee attrition and quality of life (20). The study found that low work satisfaction, a lack of prospects for advancement, and poor performance evaluations are some of the characteristics that contribute to attrition, using a dataset consisting of 1471 rows and 35 columns. Higher accuracy and precision were achieved by the random forest classifier model than by the logistic regression approach. Additionally, the research looks at cycle analysis, which offers information on the algorithms used to anticipate and stop workforce turnover. With low pay, a lot of paperwork, poor infrastructure, few opportunities for professional advancement, and a stressful work environment all negatively impacting providers' longevity and work experience, Oregon's public behavioral health system is facing a workforce crisis (21). This problem is exacerbated by ongoing underfunding and inadequate administrative support, which leaves frontline providers feeling unappreciated and dissatisfied. These problems should be addressed by policies aimed at alleviating labor shortages. With perks like social security, health insurance, and compensation, employees are essential to a company's success. But a lot of workers leave, which impacts the company's morale, productivity, and reputation. By examining firm data and using prediction algorithms such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), decision tree, decision tree C5.0, and random forest, staff attrition was assessed (22). By reducing attrition and increasing retention rates, assists businesses in enhancing profitability and coordinating objectives with staff members. The results can assist businesses in increasing profitability and coordinating their objectives with their workforce. Finding the main elements affecting contact center workers' long-term retention in Malaysia was the aim (23). Using the Delphi technique, 24 industry experts were chosen. According to the results, the most crucial elements were training, career growth, rewards, recognition, training, work-life balance, supportive management, pay, benefits, and health assistance. In order to forecast increased employee attrition from non-numerical data, a logistic and linear regression classifier was suggested (24). It highlights the advantages of logistic regression over linear regression by comparing the accuracy rates of 1470 samples of attrition datasets. In the last ten years, there has been a scarcity of music teachers in more than 30 states, and attrition has been a significant contributing cause. In Maryland, 18 percent of new music teachers leave within five years, and the greatest risk of attrition is in years

two and three. Policies recommended for mentoring, higher pay, and better resources (25). Teachers in Washington State, with hybrid and standard retirement plans, behave similarly in the years before eligibility for retirement, and so qualify as required despite differences in incentives for doing so (26). With an analysis of 4,000 employees on day 261 using logistic regression that yields 75% accuracy through the detection of 11 key attrition causes, the work concluded in its recommendation suggesting workplace condition improvement to drive employee retention and competitive edge (27, 28). A review of earlier research on employee attrition is given in Table 1, together with information on the methodology employed, datasets consulted, important variables examined, and conclusions that are pertinent to the human resources environment.

Table 1: Comparative Analysis of Employee Attrition Research

Study Reference	Features	Methodology	Dataset Details	Key Factors Identified	Key Findings	Limitations	Recommendations
(17)	HR interviews, attrition risk	HR interviews, data mining tools	HR department data	COVID-19, remote work, high-risk attrition	89% accuracy for attrition forecasting; impact of COVID-19 and remote work	Limited data diversity	Flexible work policies; improve engagement
(18)	Relationship satisfaction, overtime, environment	LR, RF, XGBoost, SVM, ANN	IBM HR data (30 features)	Relationship satisfaction, environmental contentment, overtime	Identified main drivers of attrition; talent management focus	Limited generalizability	Talent management strategies
(19)	Employee behavior, explainability	Explainable AI (XAI)	HR and behavior data	Impact on output, morale, financial security	Impact on morale and financial security; AI transparency	Limited exploration of XAI approaches	Use XAI for better decision transparency
(20)	Work satisfaction, performance	Random Forest, Logistic Regression	1471 rows, 35 columns	Low satisfaction, lack of growth, poor evaluations	Identified key attrition factors	Limited to IBM-specific dataset	Improve growth opportunities, employee evaluation
(21)	Pay, workload, support	Workforce crisis analysis	Oregon behavioral health	Low pay, workload, poor infrastructure, stress	Pay and workload affect retention	Limited to Oregon; qualitative data	Address pay, reduce workload stress
(22)	Firm-level data	KNN, SVM, Decision Tree, Random Forest	Firm-level data	Profitability, alignment, retention	Improved profitability and alignment	Firm-specific scope	Align goals with employee aspirations

(23)	Career growth, work-life balance	Delphi technique	Contact center (Malaysia)	Training, rewards, management support	Training and rewards key for retention	Small sample size; qualitative	Career development; improve work-life balance
(24)	Attrition rate comparison	Logistic Regression, Linear Regression	1470 samples (non-numeric)	Accuracy, work conditions	Logistic regression better for attrition	Limited comparison of models	Use logistic regression for non-numerical data
(25)	Early-career teacher attrition	Teacher attrition analysis	Maryland teacher data	Lack of mentorship, inadequate pay	18% quit within 5 years; suggested mentorship	Limited to Maryland early-career teachers	Implement mentorship programs, improve pay
(26)	Hybrid retirement, teacher behavior	Hybrid retirement analysis	Washington State data	Retirement eligibility, income incentives	Similar behavior near retirement eligibility	Focus on teachers; state-specific	Review incentives, retirement policies
(27)	Workplace stability, attrition causes	Logistic Regression	4,000 employees (261 days)	Workplace changes, stability	75% accuracy; identified 11 attrition factors	Short timeframe; 2015 data	Address workplace stability factors

Being able to compare clearly all the different approaches that distinct studies have undertaken in using different data sources and applying various analytical methods is worth a lot. Each entry leads to unique factors identified in the study, such as the effect of COVID-19 on remote work, job satisfaction, pay, career growth, and environmental conditions. Other limitations include data diversity and geographic specificity. Finally, each of the studies offers recommendations, ranging from flexible work policies to talent management and mentorship-all of which are actionable recommendations for organizations looking to improve retention. The approach structured here provides a useful overview for HR professionals and researchers interested in targeted strategies for reducing attrition.

Differing from a large number of previous studies focusing on predictive performance, this paper places equal significance on both prediction accuracy and explainability of models with the application of explainable AI techniques. Additionally, by employing SMOTE in class balancing, the minority class (attrition of employees) is properly represented while the model is trained. By integrating logistic regression and Random Forest with feature importance analysis, the present research brings both performance benchmarking as well as insightful decision support tools for HR managers into one comprehensive package that is not found in any existing research.

The purpose of the study is to use predictive modeling to solve the problem of staff attrition and explainable AI approaches that give a deeper insight into the reasons behind the attrition, especially in contexts that have undergone a significant shift, such as the pandemic. This study aims at developing attrition models not only forecasting with high accuracy but explaining which can aid HR practitioners in making appropriate decisions aimed at improving the retention policy. The particular objectives for this study include: determine key factors responsible for the attrition process, establish and test the predictive model to forecast the attrition rate, and propose an intervention that will reduce those factors responsible for attrition. This paper is structured along the following lines: Section. 2 discuss prior work and related studies of employee attrition analysis with predictive models, giving a general outline of the used research methods and findings from previous analyses. Section 3

describes the datasets used for this study along with methodological approach that has been adopted. Results and discussions are presented, including a description of performance of models developed and understanding derived by feature importance analysis. Recommendations for HR intervention and the future research avenues are also concluded in the last section.

Methodology

Data Set Description

The dataset for this analysis on employee attrition is obtained from Kaggle. It has HR records data from an organization, giving a very extensive view of employee demographics, performance metrics, work-related factors, compensation details, and much more. Included in the features of this dataset are Age, Job Role, Department, Monthly Income, Years at Company, and the target variable being Attrition, or if an employee has left the company, also shown in Figure 1.

This network of attributes provides insight into the various contributing factors to employee turnover. This dataset holds some numerical variables as well as categorical ones. Among them, Department, Education, JobRole, and Business Travel are the categorical ones which give qualitative data for employees. A dataset such as this is quite appropriate to be trained on machine learning models so that attrition could be predicted based on multifactorial influences behind attrition, using historical performances or satisfaction levels, or job-related metrics in a combination. The flow chart of the main steps conducted for this research study is shown in Figure 2. In order to provide an organized strategy for employee attrition study, it provides examples of the whole technique, from data collection to result interpretation.

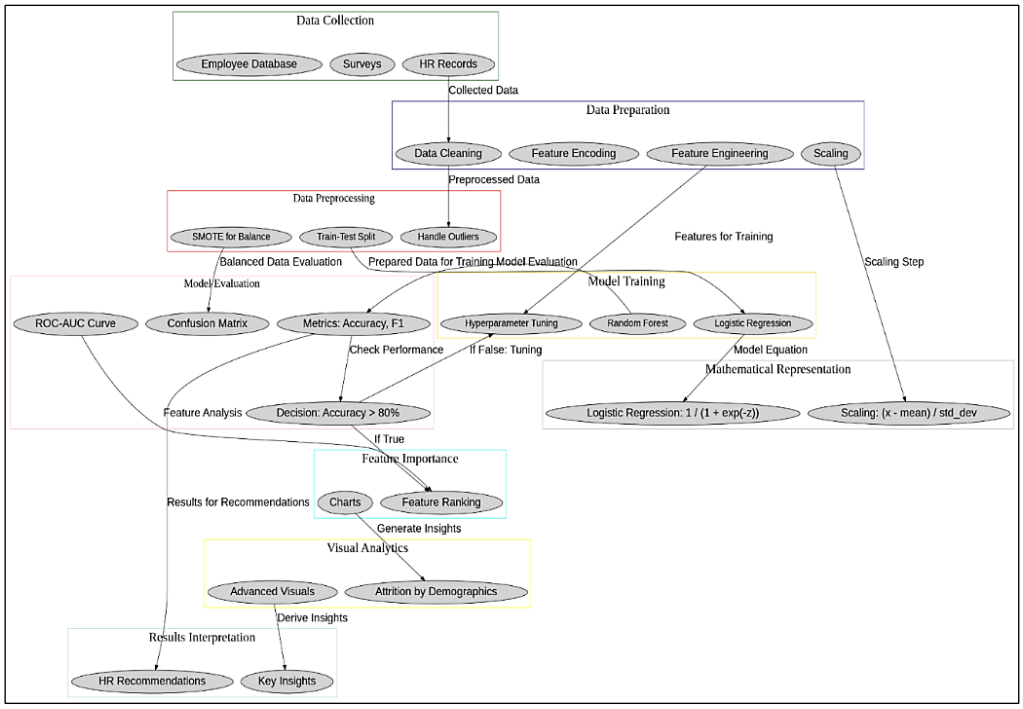

Figure 1: Flowchart of Methodology Adopted

Data Preparation

The Data Preparation phase therefore involves a series of process that ensure the dataset is cleansed, consistent, and ready for analysis as depicted in Figure 2. Initially, it deals with Missing values and Incorrect data, which may affect the input in the dataset. To begin with, missing values are usually detrimental to some models as they will deliver incorrect predictions; hence removing the rows affected or their imputation by appropriate measure, such as mean, median, will be applicable. Incorrect data entries are corrected or removed to ensure the maintenance of data integrity and enhance quality.

	Age	Attrition	BusinessTravel	DailyRate	Department	
0	41	Yes	Travel_Rarely	1102	Sales	
1	49	No	Travel_Frequently	279	Research & Development	
2	37	Yes	Travel_Rarely	1373	Research & Development	
3	33	No	Travel_Frequently	1392	Research & Development	
4	27	No	Travel_Rarely	591	Research & Development	

	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	\
0	1	2	Life Sciences	1	1	
1	8	1	Life Sciences	1	2	
2	2	2	Other	1	4	
3	3	4	Life Sciences	1	5	
4	2	1	Medical	1	7	

	...	RelationshipSatisfaction	StandardHours	StockOptionLevel	\
0	...	1	80	0	
1	...	4	80	1	
2	...	2	80	0	
3	...	3	80	0	
4	...	4	80	1	

Figure 2: Description of Dataset Collected

Another critical step in this phase was Feature Engineering. This included computing new meaningful metrics based on the already existing features. These could be further utilized to inform the model by providing new information. In this context, tenure in years since an employee was hired by the company would denote how long an employee has been at the company. Other additional metrics, like YearsSinceLastPromotion, were calculated to show where an employee was along a career trajectory and could thus provide a richer source of information about modeling attrition. In Feature Encoding, categorical variables were translated into numeric representations, suitable for use in machine learning models. For instance, JobRole and Department, being text categories, are encoded by one-hot encoding. One-hot encoding produces a column of binary values for each category, showing whether the specific value is absent or present. For example, given the feature JobRole, having three categories (such as "Manager," "Sales," "Research"), three new columns of binary values will be created for the presence or absence of the corresponding value. Scaling was applied to the numerical features so that all data was on a comparable scale by using the formula given by Equation [1]:

$$x_{scaled} = \frac{x-mean}{S.D} \qquad [1]$$

where S.D. is the norm deviation, mean is the feature's average, and x is the initial value. This transformation scales features to have a mean of 0 and a standard deviation of 1, which speeds up model convergence during training and enhances performance in general.

Data Preprocessing Steps

The Data Preprocessing phase is a combination of prepared data, leading it to be fit for the training of the model. First was Handling Outliers. These are extreme values in a data set that does not conform to the general pattern. Outliers were spotted and either treated or removed. Outliers may bias model training, and handling them enhances the robustness and reliability of the model. The next step was the balancing of the dataset through SMOTE, which stands for Synthetic Minority Oversampling Technique. Employee attrition data tends to be very imbalanced; hence the number of records showing that the employee stays is always far more than those that depict employees leaving. This could thus result in a biased model if you end up mostly with accuracy of the majority class instead of the minority one. This algorithm helps in the process of solving this problem by creating synthetic examples of the minority class to balance the data. This enhances the capability of the model in learning about attrition-related patterns well. Train-Test Split was used in order to split the data into two sets, the training set and the testing set. This split ratio is 70:30 by default, using 70% of the data for training and the other 30% for test against its performance. Split ensures a large quantity of training but a suitable amount of validation so the model would avoid overfitting—that means good training performance, whereas it underperforms at testing the new data. These stages of data preparation and preprocessing are important in preparing the dataset so that it remains consistent and makes the model strong enough to predict employee attrition. Data quality, balance in a dataset, and feature preparation are significant

factors leading to proper predictions from the machine learning models.

Model Training

This training phase started with a baseline model using Logistic Regression. Logistic Regression helps evaluate the basic performance of the model before moving to complex models. In Logistic Regression, the output is a probability that a given employee will leave the company. Logistic Regression was used as a baseline classifier because of its simplicity, interpretability, and traditional use in HR analytics for binary classification problems such as attrition prediction. Random Forest was used as an advanced ensemble model because of its strength, capability to manage feature interactions, and ability to provide feature importance scores—thus making it the best choice for extracting actionable HR insights. The comparative usage of these models enables the exploration of trade-offs between interpretability and predictive accuracy. This probability is computed with the logistic function as shown in Equation [2]:

$$P(y=1) = \frac{1}{1+e^{-z}} \qquad [2]$$

where z is a linear combination of the input characteristics calculated by Equation [3] and y=1 denotes that the employee departed the firm.

$$z = w_0 + w_1x_1 + w_2x_2 + \cdots \ldots\ldots\ldots\ldots w_nx_n \qquad [3]$$

In above equations, $w_0, w_1, w_2 \ldots\ldots w_n$ are the weights (or coefficients) learned by the model, and $x_0, x_1, x_2 \ldots\ldots x_n$ are the feature values (such as Age, Job Role, Monthly Income, etc.). The logistic function squashes the value of z between 0 and 1, providing a probability score. A more complex model was established following the baseline. The used model was Random Forest. This is an ensemble learning method that uses many decision trees combined to enhance the accuracy and reduce variance. The output is taken based on voting in the classification tasks or averaging in the regression tasks from independent prediction made by each decision tree. Since Random Forest is a combination of multiple decision trees, it does not have any single defining equation but it improves the prediction accuracy quite significantly by exploiting the diversity of trees. In order to enhance the performance of the model, hyperparameter tuning was applied. Grid search or random search were used for this purpose. To get the best outcomes, the hyperparameters were improved, such as the total number of decision trees (n_estimators) along with the maximum depth of each tree (max_depth). Hyperparameter tuning is all about finding the optimal setting that will result in maximum model accuracy without overfitting.

Model Evaluation

During Model Evaluation, different metrics were used to evaluate the performance of the learned models. The first of these metrics was Accuracy, which was calculated using Equation [4]:

$$Accuracy = \frac{TP\ (True\ Positives)+TN(True\ Negatives)}{TP\ (True\ Positives)+TN(True\ Negatives)+FP(False\ Positives)+FN\ (False\ Negatives)} \qquad [4]$$

FP, FN, TP, and TN are all determined using the matrix of error. In summarizing the prediction results, the confusion matrix illustrates how this model may differentiate across workers who are leaving and those who are staying.The F1-score, which establishes a balance between precision and recall, is another crucial statistic. This is how the F1-score is determined by Equation [5]:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad [5]$$

The ROC-AUC Curve, or Receiver Operating Characteristic - Area Under Curve, was used as evaluator of the model for its capacity to differentiate between positive classes, employee left, and negative, employee stayed for various threshold values. The AUC or Area Under Curve measures the overall performance; the value closer to 1 implies an excellent model discrimination capability. A decision point was established at which the model needed to attain greater than 80% accuracy for analysis to continue. When it failed to do so, Hyperparameter Tuning was performed to improve model performance.

Feature Importance

The Feature Importance analysis was meant to see which features are most relevant when predicting attrition from company employees. For the Random Forest model, the intrinsically estimated importance for each feature is calculated from the contribution of each feature in minimizing

uncertainty in the data. This can be performed using Gini Impurity or information gain, to mention a couple. For an internal node in the decision tree, Gini Impurity is calculated as Equation [6]:

$G = 1 - \sum_{i=1}^{n} p_i^2$ [6]

Where, pi is the fraction of samples belonging to class 'i' at the given node. Random Forest computes the average Gini decrease or information gain contributed by each feature across all decision trees to measure its importance. The features with the greatest impact on attrition for employees were the Monthly Income, Job Satisfaction, and Work-Life Balance.

Results and Discussion

This section presents the insights derived from the employee attrition analysis, detailing the performance of the models, the key features contributing to attrition, and the practical implications for HR decision-making.

Evaluation of Feature Importance

The value of feature importance analysis lies in enhancing the transparency, efficiency, and accuracy of predictive models. It informs which features are driving predictions and can therefore guide the concentration of organizational resources on impactful areas to enhance interventions and strategies. More importantly, knowing the importance of features makes it possible to build models that generalize better and hence perform well across different data sets, making them robust for real-world applications.

Feature Importance of the Random Forest model on Overtime, Marital Status, Stock Option Level, Years in Current Role, and Total Working Years as the most important in predicting employee attrition is found. Overtime is the most important feature, as shown in Figure 3, that indicates working overtime is more likely for employees to leave the organization. Marital Status also speaks to the idea that possibly family obligations or work-life balance may be the factor making the employee choose whether or not to stay or go. Other characteristics are the Stock Option Level and Years in Current Role, where it is evident that the level of monetary reward as well as career advancement factors determine whether an employee remains in the company.

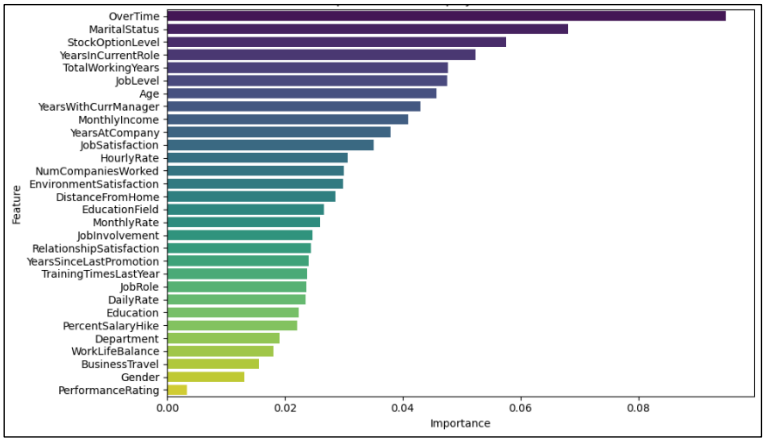

Figure 3: Feature Importance Scores for Employee Attrition

Attrition by Job Role

The organization will have a better understanding of attrition rates across all job types. Sales executives, laboratory technicians, and research scientists are among the jobs with comparatively greater turnover rates, according to Figure 4; managers, manufacturing directors, and human resources personnel left the company in comparatively fewer numbers. This would demonstrate the variations in attrition rates by job function and may even imply that stress, busyness, or dissatisfaction with a lack of prospects for growth were experienced differently in certain positions than in others. Developing retention strategies appropriate for certain high-risk roles requires this understanding. The analysis showed that workers were more likely to quit if their monthly pay was lower and they had worked for fewer years overall. This means that low pay and limited experience in the company are major factors causing turnover. In addition, younger employees appeared to be more at risk of attrition, perhaps because they want better opportunities elsewhere.

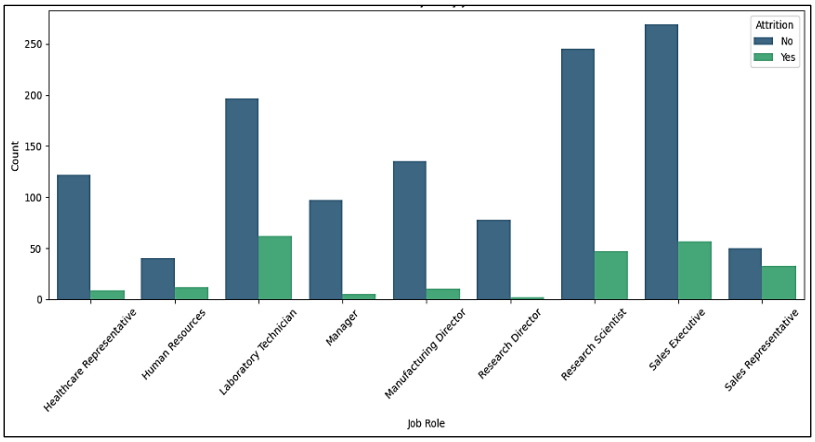

Figure 4: Attrition Analysis Based on Job Roles

Feature Importance analysis indicated that OverTime was the most important predictor of employee attrition. This emphasizes the implications of workload and stress to employees, calling for organizational policies of work-life balance, such as limiting overtime hours or offering compensatory leave. This implies that the attrition is influenced by the personal life factors, and hence, family-friendly policies or flexible work arrangements may be offered by the organizations to the employees. In a similar vein, financial incentives being significant (the Stock Option Level, for example) would imply that competitive compensation packages are critical in keeping talent on board, particularly for jobs that have a higher attrition.

Job-role attrition analysis shows some inconsistencies in the rates of turnover. High attrition rates are reported for Sales Executives, Laboratory Technicians, and Research Scientists. Such positions could be difficult or stressful for the incumbent to perform his job with ease, either due to lack of proper career growth opportunities or unsatisfying job conditions. Organizations can focus on measures specific to each role. Career development programs, distribution of workloads, and better pay packages may prove effective measures. Whereas a lower attrition rate in Managers and Human Resources indicates that such positions are probably better supported, have growth opportunities, or are less stressful and could be taken as an example to make other jobs better.

Model Performance Summary

There are several significant metrics available for evaluating the model's performance. The model's accuracy for positive predictions is shown by precision, which calculates the ratio of genuine positives to all positive forecasts. Recall assesses the model's capacity to accurately identify all true positives and the level of relevance of the occurrences it captures. The accuracy and recall harmonic mean is the F1-Score. In order to proportionately balance the two evaluations, it provides the average value with respect to unbalanced datasets. Support is the real quantity of each class's instances. The accuracy and recall values are contextualized in this way. The method summarizes true positives, true negatives, false positives, and false negatives to show how well a particular framework distinguishes between classes, namely, an employee who stayed and one who left.

Table 2: Logistic Regression Model

Metrics	Class 0 (Stayed)	Class 1 (Left)	Macro_Average	Weighted_Average	Overall
Precision_value	0.92	0.29	0.60	0.83	
Recall_value	0.77	0.57	0.67	0.75	
F1-Score_value	0.84	0.38	0.61	0.78	
Support	380	61			441
Accuracy	-	-	-	-	74.60%

Table 3: Confusion Matrix (Logistic Regression)

Predicted/Actual	Stayed (0)	Left (1)
Stayed (0)	294	86
Left (1)	26	35

Logistic Regression Model has been generated for the model and, hence, providing a prediction of employee attrition and also performance metrics in Table 2 the Class 0 that had precision of 0.92 and an F1-score of 0.84, indicating a great job in terms of detection of employees who have not left but very bad when it comes to employees that actually left: its precision was 0.29. The model achieved a general accuracy of 74.60%. Confusion Matrix of Logistic Regression in Table 3 shows the actual versus predicted outcomes. In this, the model accurately predicts 294 stayers and 35 leavers. There are 86 false positives, that is, stayers predicted as leavers, and 26 false negatives, that is, leavers predicted as stayers.

The Random Forest Model offers statistics for a more sophisticated model, according to Table 4. The Random Forest accuracy was 85.71% whereas the Logistic Regression was lower. Precision Class 0 is 0.88, but it does poorly with Class 1 at 0.46 precision and 0.18 recall, unable to find who has actually left. According to the presented confusion matrix in Table 5, this model would accurately predict 367 people to stay and 11 people to leave. There are fewer false positives than with logistic regression, however the model has a flaw with 50 false negatives, which would indicate incorrect attrition.

Table 4: Random Forest Model

Metrics	Class 0 (Stayed)	Class 1 (Left)	Macro Avg	Weighted Avg	Overall
Precision_value	0.88	0.46	0.67	0.82	
Recall_value	0.97	0.18	0.57	0.86	
F1-Score_value	0.92	0.26	0.59	0.83	
Support	380	61			441
Accuracy	-	-	-	-	85.71%

Table 5: Confusion Matrix (Random Forest)

Predicted/Actual	Stayed (0)	Left (1)
Stayed (0)	367	13
Left (1)	50	11

Based on the findings, several HR interventions are suggested to deal with the problems of employee attrition. For example, reduction in overtime hours and a healthy work-life balance would be able to deal with one of the major causes of turnover. Career development programs and financial incentives like stock options should be introduced to encourage employees to stay. There should be special attention to those roles that fall into the higher range of attrition, such as Sales Executives and Laboratory Technicians, in specific development and support initiatives. In addition, programs like mentorship might also be a good approach with younger employees as well as those with short tenures, encouraging engagement and commitment from their very first career step.

Conclusion

The analysis for employee attrition using various machine learning models reveals important factors that influence the desire of employees to stay at or leave an organization. Among those factors, it was overtime, marital status, stock option level, and job role that ranked as significant determinants, implying that both work-related stressors and financial incentives play an important role within

the phenomenon of employee attrition. The Random Forest model was more accurate than the Logistic Regression model and found the patterns of attrition. The conclusions drawn from this study give actionable recommendations for HR professionals to develop targeted strategies that reduce attrition. Overtime-related issues and competitive compensation packages as well as career development and mentorship can enhance employee satisfaction and retention. This study emphasizes the benefit of HR practices that work in accordance with work-life balance and employee well-being to create a stable workforce. Other features, such as performing time-series analysis of patterns of attrition, sentiment monitoring of employees, or complex machine learning algorithms, can be incorporated to strengthen the conclusions drawn from the study. Explanatory AI will assist HR managers in taking rational decisions since it will make better model interpretability possible. Due to the very unique data set that is used in the study, and unequal class distribution with a focus on only the quantitative characteristics, which is unable to represent all the qualities of employee engagement and happiness, such as an interpersonal relationship or workplace culture, it may not be generalized further. Future research can explore integrating deep learning models and time-series employee behavior data for dynamic attrition prediction. Additionally, leveraging sentiment analysis from employee feedback, or incorporating qualitative factors like organizational culture, could further enhance predictive capability. Expanding the dataset across multiple companies and industries would also help generalize findings.

Abbreviations

ANN: Artificial Neural Network, HR: Human Resource, KNN: K-Nearest Neighbors, LR: Logistic Regression, RF: Random Forest, SVM: Support Vector Machine, XAI: Explainable AI.

Acknowledgement

We extend our heartfelt thanks to the faculty and staff of the participating institutions for their invaluable support and contributions to this research. Special gratitude is directed to the reviewers and colleagues who provided insightful feedback and guidance that greatly enhanced the quality of this manuscript.

Author Contributions

Mustafizul Haque: Conceptualization, Literature Review, drafting the manuscript, Tejasvini Alok Paralkar: Methodology, Data Analysis, Review, Editing, Sudhir Rajguru: Literature Review, Proofreading, Adheer A Goyal: Data Collection, Review, Editing, Final Approval, Tanaya Patil: Visualization, Literature Review, Kamal Upreti: Methodology, Data Analysis, Review, Editing.

Conflict of Interest

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article. All authors declare no conflict of interest for this article.

Ethics Approval

No Participation of humans take place in this implementation process.

Funding

No Funding support is provided for this paper publication.

References

1. Al-Alawi AI, Ghanem YA. Predicting Employee Attrition Using Machine Learning: A Systematic Literature Review. In2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS). IEEE. 2024 Jan 28:526-530. https://doi.org/10.1109/icetsis61505.2024.10459451
2. Md Atikur RA, Md Sayed UD, Wasib Bin LA. Effects of training and development, Organizational culture, Job satisfaction, and career development on employee retention in commercial banks in Bangladesh. The Journal of Asian Finance, Economics and Business. 2023;10(2):91-7.
3. Phakdeechanuan K, Kellett U, Henderson S, Corones-Watkins K, Saito A, Thiangchanya P. Addressing Registered Nurse Retention and Attrition in Thailand Hospitals: An Integrative Review. Asia Pacific Journal of Public Health. 2025 Jan;37(1):17-29.
4. Bhargav S, Mehra N. Study of Employee Attrition in Business Process Outsourcing Companies in India. International Journal of Research in Social Sciences. 2018 Jan1;8(9):348–58.
5. Kapoor M, Chowdhury JK. Employee attrition & retention strategy in BPO sector: A review of literature. ZENITH International Journal of Multidisciplinary Research. 2018;8(11):79-83.
6. Irabor IE, Okolie UC. A review of employees' job satisfaction and its affect on their retention. Annals of Spiru Haret University. Economic Series. 2019 Jun 28;19(2):93-114.
7. Kumar TS, Kavitha M. Employee retention- A real time challenges in Indian IT sector - Review paper.

Asian Journal of Multidimensional Research. 2018 Jan 1;7(6):83–95.
8. Ravesangar K, Narayanan S. Adoption of HR analytics to enhance employee retention in the workplace: A review. Human Resources Management and Services. 2024 Aug 12;6(3):3481.
9. Haque F. Retention of tech employees in India: lessons from the extant literature. The Learning Organization. 2024 Jan 8;31(4):585–629.
10. Nandal M, Grover V, Sahu D, Dogra M. Employee Attrition: Analysis of Data Driven Models. EAI Endorsed Trans. Internet Things. 2024;10:1-10.
11. Alkaabi A, Alghizzawi M, Daoud MK, Ezmigna I. Factors affecting employee turnover intention: An integrative perspective. InThe AI Revolution: Driving Business Innovation and Research. Cham: Springer Nature Switzerland. 2024 Jun 18;2: 737-748.
12. Usha PM, Balaji NV. A comparative study on machine learning algorithms for employee attrition prediction. IOP Conference Series Materials Science and Engineering. 2021 Feb 1;1085(1):012029.
13. Seelam SR, Kumar KH, Supritha MS, Gnaneswar G, Reddy VV. Comparative study of predictive models to estimate employee attrition. In2022 7th International conference on communication and electronics systems (ICCES). 2022 Jun 22:1602-1607. IEEE. https://doi.org/10.1109/icces54183.2022.9835964
14. Nigam J A S, Barker RM, Cunningham TR, Swanson NG, Chosewood LC. Vital Signs: Health Worker–Perceived Working Conditions and Symptoms of Poor Mental Health — Quality of Worklife Survey, United States, 2018–2022. MMWR Morbidity and Mortality Weekly Report. 2023 Oct 24;72(44):1197–1205.
15. Wu TJ, Yuan KS, Yen DC. Leader-member exchange, turnover intention and presenteeism– the moderated mediating effect of perceived organizational support. Current Psychology. 2021 May 13;42(6):4873–84.
16. Gelencsér M, Szabó-Szentgróti G, Kőmüves ZS, Hollósy-Vadász G. The Holistic Model of Labour Retention: The Impact of Workplace Wellbeing Factors on Employee Retention. Administrative Sciences. 2023 May 1;13(5):121.
17. Jung J, Kim BJ, Kim MJ. The effect of unstable job on employee's turnover intention: The importance of coaching leadership. Frontiers in Public Health. 2023 Mar 16;11:1068293.
18. Mozaffari F, Rahimi M, Yazdani H, Sohrabi B. Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data. Benchmarking an International Journal. 2022 Dec 14;30(10):4140–73.
19. Chung D, Yun J, Lee J, Jeon Y. Predictive model of employee attrition based on stacking ensemble learning. Expert Systems With Applications. 2022 Nov 30;215:119364.
20. Díaz GM, Hernández JJG, Salvador JLG. Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making. Mathematics. 2023 Nov 17;11(22):4677.
21. Sharma S, Sharma K. Analyzing Employee's Attrition and Turnover at Organization Using Machine learning Technique. In2023 3rd International Conference on Intelligent Technologies (CONIT).IEEE. 2023 Jun 23:1-7. https://doi.org/10.1109/conit59222.2023.10205676
22. Hallett E, Simeon E, Amba V, Howington D, McConnell KJ, Zhu JM. Factors Influencing Turnover and Attrition in the Public Behavioral Health System Workforce: Qualitative Study. Psychiatric Services. 2023 Jun 30;75(1):55–63.
23. Jain R, Nayyar A. Predicting employee attrition using xgboost machine learning approach. In2018 international conference on system modeling & advancement in research trends (smart).IEEE. 2018 Nov 23:113-120. https://ieeexplore.ieee.org/abstract/document/8746940
24. Subramaniam SH, Wider W, Tanucan JC, Yew Lim K, Jiang L, Prompanyo M. Key factors influencing long-term retention among Contact Centre employee in Malaysia: a Delphi method study. Cogent Business and Management. 2024 Dec 31;11(1):2370444.
25. Khekare G, Balaji K, Arora M, Tirpude RR, Chahar B, Bodhankar A. Logistic and linear regression classifier based increasing accuracy of non-numerical data for prediction of enhanced employee attrition. In2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE).IEEE. 2023 May 12 :758-761. IEEE. https://doi.org/10.1109/icacite57410.2023.10183226
26. Miller DS. The retention and attrition of early-career music teachers: A survival analysis. Arts Education Policy Review. 2025 Apr 3;126(2):82-98.
27. Goldhaber D, Grout C, Holden KL, McGee JB. Evidence on the Relationship between Pension-Driven Financial Incentives and Late-Career Attrition: Implications for Pension Reform. ILR Review. 2024 Jan 10;77(2):175–98.
28. Setiawan I, Suprihanto S, Nugraha AC, Hutahaean J. HR analytics: Employee attrition analysis using logistic regression. IOP Conference Series Materials Science and Engineering. 2020 Apr 1;830(3):032001.