Applied Science, Engineering and Management Bulletin [ASEMB]
ISSN: 3049-3005
Vol. 3 No. 02(Apr-June) (2026) |Pg:79-85

Predicting Employee Attrition Using Ensemble Machine Learning Techniques: A Comparative Study

Dr.S.Kevin Andrews[1], Dr.Praveen B M[2], G.Anandhi[3], Mr.Jayakumar[4]
[1]Post Doctoral fellow, Srinivas University, Mukka, Mangalore
Kevin.mca@drmgrdu.ac.in
[2]Srinivas University, Mukka, Mangalore
[3]Assistant Professor, Dept of Computer Applications, Dr.MGR Educational and Research Institute, Chennai- 95
anandhi.mca@drmgrdu.ac.in
[4]Research Scholar, Department of Computer Applications, Dr.MGR Educational and Research Institute, Chennai-95

Abstract: *One of the most important problems to solve for any organization is employee attrition because of the effect on workforce stability and loss of knowledge and experience. Historical HR analytics solutions are likely to be reactionary and might not be able to identify complex relationships in employee data. Based on this work, a soft voting scheme is proposed for the ensemble prediction of employee attrition with high accuracy. Thee research consists of the IBM HR Analytics dataset, containing 1,470 employee records and 35 features, from the Kaggle database. Synthetic Minority Over-Sampling Technique (SMOTE) and Random Over-Sampling (ROS) methods are used to tackle the class imbalance problem. The proposed ensemble is a combination of Extreme Gradient Boosting (XGBoost), Random Forest and calibrated Logistic Regression classifiers. The experimental results indicate that the accuracy of the model is 97.72% and the F1 score is 97.74%. The results prove that data-driven predictive systems can play a key role in implementing proactive employee retention strategies.*

Keywords:*Ensemble Learning, Random Over-Sampling, Employee Attrition, Predictive HR Analytics, SMOTE, Soft Voting*

1. Introduction

In today's business world, human capital is vital to sustaining competitiveness and helping to facilitate ongoing innovation within organizations. Regular staff changes can cause disruptions in the business, employee productivity losses, and hiring and training costs. It is estimated that it costs 1.5 to 2 times an employee's annual salary to replace a skilled employee (Baruaet al., 2025). The traditional HR practices are largely based on historical reports and exit interviews, which are reactive rather than proactive. This study, based on these drawbacks, proposes a comparative study of ensemble machine learning techniques to predict employee attrition prior to their departure from the organization. Predictive analytics can help companies analyze the likelihood of employee attrition, uncover issues in the workplace and create specific retention strategies to help minimize employee turnover and boost the retention of valuable employees.

2. Literature Review

Researchers have moved beyond simple statistical methods to more sophisticated machine learning techniques to enhance prediction accuracy in the workforce analytics field.

2.1 Class Imbalance in Human Resource Datasets

The data in employees are naturally imbalanced—most employees stay in a company over a period of time. Machine learning models based on these data sets can, for example, be very accurate but poor in predicting employees who are at risk of resignation(Bandaru et al., 2025). For this reason, oversampling has been extensively used. The imbalance can be addressed by creating new instances of the minority class, e.g., using Synthetic Minority Over-sampling Technique (SMOTE) or Adaptive Synthetic Sampling (ADASYN) (Eom&Byeon, 2023). These methods help increase the sensitivity of the models and lower the chances of missing significant patterns of attrition.

2.2 Performance of Standalone Tree-Based Classifiers

Structured HR datasets are typically structured and have been widely applied to tree-based algorithms. Random Forest, which is a committee of multiple decision trees, using different subsets of features has been consistently

effective in prediction, and frequently beat traditional algorithms like Logistic Regression (Abdulhafedh, 2022). Likewise, extreme gradient boosting (XGBoost) and LightGBM are popular boosters due to their efficient optimization and good predictive performance (Sibindi et al., 2023). While good results can be obtained using these stand-alone models, they can also overfit when trained on noisy employee samples, necessitating the use of more sophisticated ensemble models.

2.3 Ensemble Architectures and Stacking Approaches

Researchers have been increasingly turning to ensemble learning techniques to enhance the prediction performance. The stacking and soft voting techniques combine the advantages of various machine learning algorithms (Soni et al., 2026). It has been demonstrated that stacking algorithms with XGBoost and Random Forest as base learners and Logistic Regression as a meta-classifier can attain high prediction accuracy (Ali et al., 2022). The soft voting ensemble is an ensemble of models that work by averaging the probabilities that the models predict, and thus it can be used to reduce the generalization error and increase the stability of the results. This can be helpful in capturing the complex and non-linear aspects of employee behavior and turn-over.

2.4 Algorithmic Interpretability and Explainable AI

A drawback to using machine learning in HR is the lack of transparency for sophisticated predictive models. Human resource managers need explanations that are easy to understand before they act based on the outputs from models. More recently, Explainable Artificial Intelligence (XAI) approaches, like SHapley Additive exPlanations (SHAP) or Local Interpretable Model-Agnostic Explanations (LIME), have been implemented to enhance transparency (Wang, 2024). These techniques can be used to determine factors that are the most significant for employee turnover. Overtime, lower pay, poor work-life balance, and lacking stock ownership have been identified as factors that drive employee turnover, informing an organization's ability to make informed decisions regarding employee retention.

3. Methodology

The proposed framework has a structured data-processing pipeline that collects, prepares, balances and classifies employee data. This study utilizes the data set of IBM HR Analytics that comprises 1470 employee records and 35 features related to employee demographics, organization, and compensation (Miracle 'Nifise, 2025).

Data pre-processing:Missing or inconsistent values are filled in by median imputation. Various techniques are employed for processing categorical variables, such as label encoding and one-hot encoding, and numerical variables are processed using normalization techniques to minimize the impact of scaling differences on the model.

One of the major problems in the dataset is the imbalance of the class, in which around 84% of employees have the class “No Attrition” and 16% have the class “Attrition”. Therefore, in order to overcome this problem, the framework uses SMOTE and Random Over-Sampling (ROS). With the SMOTE process, new samples of the minority class are created as follows:

$$x_{new} = x_i + \lambda(x_k - x_i) \quad \dots (i)$$

A minority-class sample in this equation (equation i) is denoted by $x_i \in X_{minority}$, its k-nearest minority-class neighbor is denoted by x_k and λ is a random value from a uniform distribution. This approach is conducted within the feature space to provide realistic synthetic observations that help in balancing the dataset without making a duplicate of existing observations.

An 80:20 split of the balanced dataset into training and testing sets is performed by using a stratified split. A stratified split is used to split the balanced dataset into 80% training and 20% testing sets. Moreover, fivefold cross validation techniques are used for increasing reliability and to avoid overfitting (Yates et al., 2023). The three machine learning models trained in the framework are Random Forest, XGBoost and Logistic Regression.

The study tries to combine these classifiers by using the soft-voting ensemble approach without relying on a single classifier. The ensemble combines the prediction probabilities of the different models, giving weights based on their strengths, and yields calibrated predictions of the attrition risk scores.

The soft voting probability for a target class c is:

$$P(y = c|x) = \frac{1}{\sum_{m-1}^{M} w_m} \sum_{m-1}^{M} w_m P_m(y = c|x) \quad \dots (ii)$$

In equation (ii), each classifier m is assigned a weight w_m based on its performance on the validation set, and $P_m(y = c|x)$ is the probability that the model predicts for class c. Total number of the constituent classifier is denoted by M. Combining boosting, bagging, and linear classification methods in a single ensemble, this approach will minimize both the variance and the bias and produce a practical and reliable employee attrition prediction system. The final model is a tool for organizations to use for predicting their turnover risk and making decisions about retention planning based on data.

4. Analysis and Interpretation

The first step in evaluating the new predictive framework was to perform an exploratory analysis on the 1,470 employee records included in the IBM HR Analytics data set. The goal was to determine the primary reasons for employee turnover that employees are willing to report. The descriptive summary of certain important numerical and categorical variables is given in Table I and the difference between employees who stayed in the organization and those who were lost are highlighted.

Table 1: Descriptive Statistics of Key Predictors of Employee Attrition

Organizational Feature	Attrition = Yes	Attrition = No	Demographic/Structural Correlation
Mean Age (Years)	33.6	37.5	Younger employees show higher turnover risk
Mean Monthly Income	$4,787	$7,332	Strong negative correlation with attrition
Commute Distance (km)	10.6	8.9	Longer travel distance increases exit rates
Overtime Exposure	53.50%	23.40%	Overtime is a critical trigger for resignation
Mean Years at Company	5.1	7.3	Retention increases with employee seniority

As shown in Table 1, there are distinct differences identified between those employees who choose to stay with the company and those who leave. Those who were leaving the company tended to be younger, paid less, worked further away from home, and were more likely to do overtime work. Additional analysis indicates that younger workers and those who had worked with the company for less time were more likely to leave. This trend could be related to restricted career progression and professional development opportunities. In general, the results indicate that compensation, workload, and commuting demands among the other factors are the most significant aspects that affect employee attrition.To investigate the impact of the class imbalance, the machine learning models were first trained with the original dataset without any class balancing method. Under this condition the following results are obtained in Table 2.

Table 2: Model performances in the Unbalanced Data Configuration

Machine Learning Model	Classification Accuracy	Precision	Recall (Attrition)	F1-Score	ROC-AUC
Logistic Regression	0.864	0.721	0.354	0.475	0.812
Random Forest	0.857	0.8	0.169	0.279	0.801
XGBoost (Tuned)	0.871	0.783	0.316	0.451	0.824
Proposed Soft Voting	0.884	0.812	0.38	0.518	0.843

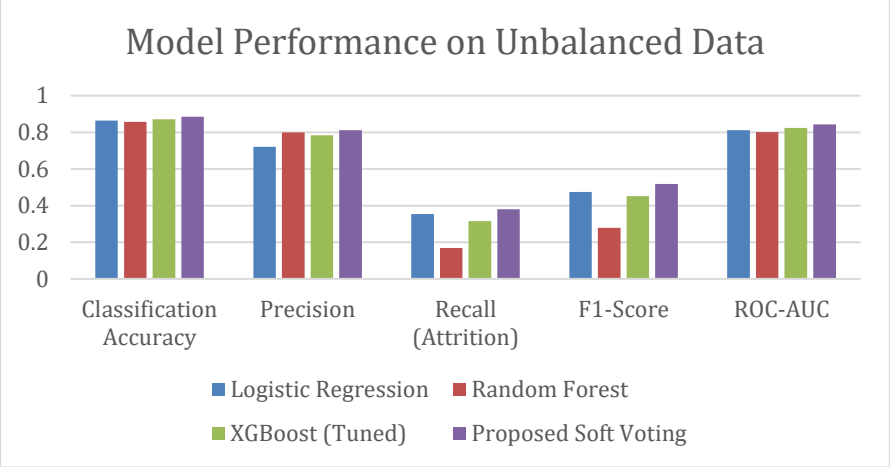

Figure 1: Comparative Performance of Models on Unbalanced Data

The results presented in Figure 1 emphasize an interesting classification problem. The relatively high accuracy values obtained for the models (85.7% to 88.4%) did not lead to a high level of recognition accuracy for those who really left the company. An example of this is that the model Random Forest had a recall of just 16.9% for the class attrition. This is because the majority of records are of the "No Attrition" type, which make up about 84% of the data set. This causes the models to be biased towards the most common class while training. Many employees who were likely to leave were falsely categorized as retained employees, due to the low recall values. As a result, although they were reasonably accurate, they were not appropriate for good employee retention planning.

Various re-sampling techniques were added to the framework for enhancing predictive performance. The performance of the soft-voting ensemble with different techniques for handling class imbalance is contrasted, in Table 3.

Table 3: The Performance Metrics for The Different Alternative Class Imbalance Remediation Schemes

Resampling Technique	Accuracy	Precision	Recall (Attrition)	F1-Score	ROC-AUC
SMOTE	0.9417	0.9412	0.9406	0.9409	0.9787
ADASYN	0.9416	0.941	0.9408	0.9409	0.9794
SMOTEENN	0.9656	0.9642	0.9803	0.9722	0.9907
SMOTETomek	0.9405	0.9392	0.9384	0.9388	0.9795
Random Over-Sampling (ROS)	0.9772	0.977	0.9778	0.9774	0.9951

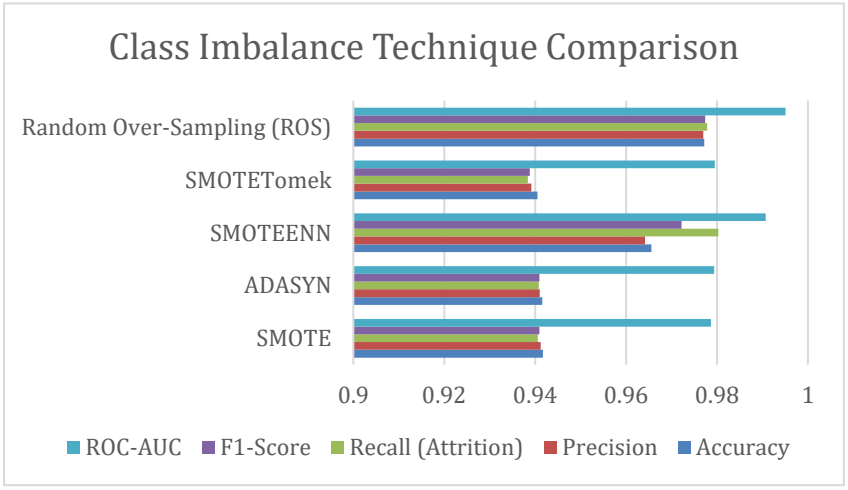

Figure 2: Impact of Class Imbalance Handling Techniques on Model Performance

As can be seen in the results in Figure 2, the balanced dataset leads to a better classification performance. The best result was achieved by the combination of the proposed soft-voting ensemble and Random Over-Sampling (ROS) technique among all the tested techniques. The model achieved an accuracy of 97.72%, an F1-score of 97.74%, and a ROC-AUC value of 0.9951. The SMOTEENN technique also performed well with an F1 score of 97.22%. But ROS yielded slightly better results, and it made a more stable and simpler data balancing process. Models suitability for practical use in the HR field was also checked from the point of view of computational efficiency. Table 4 summarizes the training time, memory usage and statistical significance results for the different resampling methods.

Table 4: Computational Efficiency and Statistical Significance Metrics

Resampling Algorithm	Execution Time (s)	Memory Consumed (MB)	ROS Comparison (p-Value)	Statistical Significance
SMOTE	11.02	1.25	0.437	Non-Significant
ADASYN	11.09	1.11	0.355	Non-Significant
SMOTEENN	10.42	1.21	0.999	Non-Significant
SMOTETomek	9.59	1.13	0.119	Non-Significant
Random Over-Sampling (ROS)	8.9	0.98	Reference Baseline	Reference Baseline

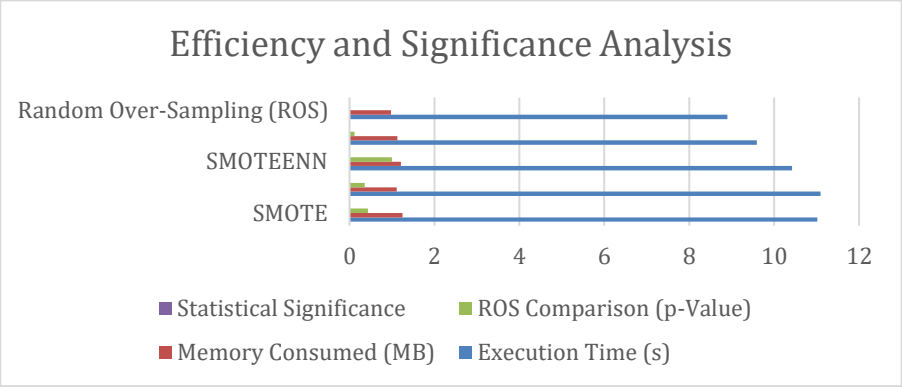

Figure 3: Efficiency and Statistical Validation of Predictive Models

As can be seen in Figure 3, the method of ROS gave the lowest computational requirements when compared with the other methods tested. The training process took 8.90 seconds for the model and used up 0.98 MB of memory. A paired t test was performed to compare the performance of alternative resampling methods to ROS. This resulted in P-values > 0.05 and thus the observed performance difference was not statistically significant. Based on these results, it is evident that the ensemble based on ROS achieved the best accuracy and efficiency, which is beneficial for its application in the organization.

5. Discussion

The findings from the experiment indicate that predictive HR analytics can enable organizations to transition from reactive HR decision making to a more proactive approach to workforce management. Random Over-Sampling in combination with a soft voting ensemble gave very stable prediction results and were able to solve the limitations reported in previous studies.

The framework proposed was able to attain excellent performance and this is attributed to the use of three different machine learning algorithms—Random Forest, XGBoost, and Logistic Regression. When it comes to predicting, each model offers something different.

- ***Random Forest:***Reduce variance – bagging,
- ***XGBoost:***Improve learning – boosting,
- ***Logistic Regression:***Provide a consistent linear baseline.

The ensemble approach can be used together to explain the complex interactions between employee characteristics like job satisfaction, overtime working, wages and organizational involvement.

The comparison of resampling methods also has a lot of good practice learning. The outcomes revealed that the performance of an easy-to-use technique such as the Random Over-Sampling method is similar to or even higher than more complex techniques like SMOTEENN and ADASYN. This discovery is especially significant in organizations where the amount of training time, memory usage, and system efficiency are important factors. This leads to improved prediction quality with minimal computational resources and renders ROS a viable way to utilize it in long-term HR analytics systems.

This study has the disadvantage of using a secondary dataset that is cross-sectional. This may not reflect over time all the changes in the behaviour of employees and/or in the broader economy. This limitation, however, is not a substantial one when considering the overall results, as the variables that were most significant in the study – compensation, overtime, and job involvement – have all been found to be important factors affecting employee turnover in a wide range of industries. For future implementation, the accuracy of the prediction can be further increased by feeding real-time organizational information to the model like employee engagement surveys, performance evaluations, or workplace feedback.

6. Conclusion and Future Directions

In this study, an ensemble of ML models has been presented for predicting employee attrition and compared with each other. The proposed model outperformed all other models with an accuracy of 97.72%, F1 score of 97.74%, and ROC-AUC of 0.9951, due to the combination of Random Over-Sampling to handle class imbalance and the use of multiple classifiers in a soft voting framework. The results show that the use of different machine learning models can significantly boost the prediction accuracy when compared to single models. Consequently, the framework provides an organization with a reliable solution to uncover employees that might be considered at risk of leaving, and it can help to facilitate the creation of proactive retention strategies.

There are a number of future research opportunities. First, graph neural networks can be considered as a means of understanding how employee turnovers can spread through teams and departments and how to better represent the relationships between employees in the workplace. Second, future models could also use real-time employee feedback and sentiment data to detect short-term shifts in employee behaviour. Finally, the framework could be coupled with recommendation systems to suggest personalized retention approaches in light of the risk profile of each employee.This would further solidify the use of predictive analytics in HRM.

References

1. Barua, R., Ghosh, S. K., Rahman, M. N., Biswas, K. C., & Haque, M. S. (2025). Optimizing Software Engineering Careers: Hiring, Retention, and Workforce Development. *Journal Of Creative Writing (ISSN-2410-6259)*, *9*(1), 49-65.https://doi.org/10.70771/JOCW.148
2. Bandaru, S., Kishore, A., Soni, M., Pareek, P., Tewatia, N., &Dayama, R. S. (2025). Strengthening Data Confidentiality with Next-Generation Cloud Security in Multi-Tenant Environments. *2025 IEEE International Conference on Communication Networks and Computing (CNC)*, 1352–1358. https://doi.org/10.1109/cnc68716.2025.11484548
3. Eom, G., &Byeon, H. (2023). Searching for optimal oversampling to process imbalanced data: Generative adversarial networks and synthetic minority over-sampling technique. *Mathematics*, *11*(16), 3605.https://doi.org/10.3390/math11163605
4. Abdulhafedh, A. (2022). Comparison between common statistical modeling techniques used in research, including: discriminant analysis vs logistic regression, ridge regression vs LASSO, and decision tree vs random forest. *Open Access Library Journal*, *9*(2), 1-19.https://doi.org/10.4236/oalib.1108414
5. Sibindi, R., Mwangi, R. W., &Waititu, A. G. (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. *Engineering Reports*, *5*(4), e12599.https://doi.org/10.1002/eng2.12599
6. Soni, M., Singh, T., Wawage, P., Tonde, D., Borkar, P., &Nadaf, J. S. (2026, March). Compact Edge-AI Architecture for Real-Time Decision Making in Smart Devices. In *2026 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 1-6). IEEE. https://doi.org/10.1109/ESCI68015.2026.11493408

7. Ali, M., Haider, M. N., Lashari, S. A., Sharif, W., Khan, A., & Ramli, D. A. (2022). Stacking classifier with random forest functioning as a meta classifier for diabetes diseases classification. *Procedia Computer Science*, *207*, 3459-3468.https://doi.org/10.1016/j.procs.2022.09.404
8. Wang, Y. (2024). A comparative analysis of model agnostic techniques for explainable artificial intelligence. *Research Reports on Computer Science*, 25-33.https://ojs.wiserpub.com/index.php/RRCS/article/download/4750/2484
9. Miracle 'Nifise. (2025). *Predicting Employee Turnover at IBM*. Kaggle.com. https://www.kaggle.com/datasets/miraclenifise/hr-employee-attrition-datasets
10. Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross validation for model selection: a review with examples from ecology. *Ecological monographs*, *93*(1), e1557.https://doi.org/10.1002/ecm.1557