

Data-Driven Workforce Analytics Using Supervised Learning Models in HRM

Dr Melanie Lourens¹, Dr. G Yugesh Kumar², Deepa K³, Dr. Satish Jha⁴

¹Deputy Dean Faculty of Management Sciences, Durban University of Technology
South Africa, Email: melaniel@dut.ac.za

²Assistant Professor, Department of Management, KPR College of Arts Science and Research, Bharathiar
University, Email: gyugeshkmr@gmail.com

³Research Scholar, Sri Ramakrishna College of Arts and Science, Coimbatore, Management, Bharathiar
University, Email: deepamphilhr@gmail.com

⁴Dean, Faculty of Hotel Management and Catering Technology, Parul University, Waghodia, Vadodara,
Gujarat, 391760, India, Email: satish.jha144400@paruluniversity.ac.in

Abstract.

Attrition by employees poses several challenges in workforce management, recruitment, knowledge management, and cost aspects of organizations. Workforce analytics can be used in HRM to understand trends from the collected employee data and determine the employment profile that is prone to attrition. In this paper, four supervised learning models—logistic regression, decision tree, random forest, and gradient boosting—are applied on the IBM HR analytics dataset on employee attrition and performance. The IBM dataset has 1,470 records and 35 variables with attrition being in the form of binary. Attrition among 237 employees (16.12%) was observed using exploratory analysis. More attrition cases were observed among employees working overtime compared to those not working overtime. Sales Representatives had the highest role attrition rate. Excluding the target, employee ID, and constant variables, 30 predictors were included. For evaluation, a stratified train-test split of 80% to 20%, as well as five-fold stratified cross-validation on the training set, was applied. The highest test F1-score (0.451) and recall (0.638) were recorded for Logistic Regression, while the highest accuracy (0.861) and ROC-AUC (0.805) were scored by Gradient Boosting. Overtime was the most important predictor using permutation importance, as identified for the chosen Logistic Regression model. It demonstrates the importance of evaluating the performance of the minority class in addition to accuracy. Since the data used in the analysis are synthetic, the results need to be interpreted as an exercise.

Keywords: workforce analytics, human resource management, employee attrition, supervised learning, Logistic Regression, Random Forest, explainable analytics

1. Introduction

HRM also relies more on computerized systems for functions such as recruitment, retention, training, and planning of the work force. Information about employees might be available on various issues related to their performance like role, tenure, salary, satisfaction, behaviour at work, and their career development. Analysis by experts will aid in detecting patterns and assessing programs aimed at retaining staff members, connecting reporting to planning. Employee turnover is important as the departures have potential to cause capacity disruption, loss of organizational knowledge, and the need for recruitment and induction of new workers. There are many factors affecting this phenomenon and it cannot be attributed to just one single measure. Supervised algorithms predict the relationship between features of employees and outcomes. There have been developments in research on machine learning for predicting employee turnover. A survey of 52 articles revealed that supervised approaches predominated in such research and highlighted salary and overtime among other significant predictive factors [1]. Organizational context and analysis-to-decision process are also crucial for workforce analytics [2]. The current study compares various modelling techniques, evaluates minority-class metrics, and cautiously interprets feature importance.

The question this paper seeks to answer is:

(1) What are the trends in employee attrition?

- (2) How do the four models differ from each other in their predictive performance on the hold-out dataset?
- (3) What features influence the F1-score of the selected model?

The contribution of the paper is a baseline analysis focused on class imbalance and data peculiarities.

2. Literature Review

There have been studies on statistical learning, decision trees, ensemble methods, and neural networks. There was a study by Alsheref et al., which introduced an ensemble method for predicting attrition [3]. There are comparative studies showing that the ranking of models depends on the metric used for evaluation: high accuracy does not necessarily imply good prediction for minority-class cases [4].

Job conditions, pay, job tenure, job satisfaction, and working conditions are frequent variables studied. In a recent review, there were studies which used Random Forest algorithm and noted the presence of job satisfaction [5]. However, importance of a variable is not equivalent to causality; the variable could be correlated to other variables as well.

The relevance of HR analytics rests on whether the findings can be trusted and put into use. Recent studies on HR data show the need to balance predictive performance and explainability [6]. The latter aspect is critical in employment-related settings, where lack of explainability of the findings may affect people's career prospects or personal life.

New approaches have been developed that apply clustering, data augmentation, and optimal neural networks for forecasting employee turnover [7]. Such techniques are more complex and need proper validation. A comparison of Logistic Regression with tree-based and boosting models on a small dataset is reasonable; success on one benchmark does not mean success on others. Furthermore, turnover prediction is considered from the perspective of resource conservation and workforce stability [8].

3. Methodology

This part talks about the methodology that was used to study employee attrition with the help of supervised machine learning models. The research employs a rigorous methodology which includes steps like preparation of dataset, exploratory data analysis, feature selection, modeling and performance evaluation. Classification algorithms such as Logistic Regression, Decision Tree, Random Forest and Gradient Boosting have been implemented and compared for the purpose of determining their ability to predict employee attrition. Suitable preprocessing methods along with stratified data splitting and cross-validation have been applied for improving the robustness of the performance evaluation process. Several performance measures are being used since the dataset has more number of employees who stayed than the number of those who left.

3.1 Dataset

Data analysis is done using the dataset provided by IBM in an Excel spreadsheet format on 'HR Analytics Employee Attrition and Performance'. The number of observations is 1,470 and the number of variables is 35. Attrition is the dependent variable with Yes or No denoting an employee being labelled as having left (Attrition = Yes) or still working with the company (Attrition = No).

3.2 Data Preparation and Exploratory Analysis

Target was encoded as a binary variable: yes = 1 and no = 0. The data set was examined for any missing data and there was none in the worksheet imported into the workbook. EmployeeNumber was dropped as it is not a meaningful predictor but an identifier. EmployeeCount, Over18, and StandardHours were dropped since they were constant in this particular dataset. Target variable was also dropped from predictors, thus, there were 30 input variables.

Class Balance Analysis provided summary of class balance and comparison of attrition depending on overtime and job role as well as income by attrition status. Rates were calculated for each group separately. It is descriptive statistics without controlling for any confounders.

3.3 Model Development

Four supervised classifiers were evaluated: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Logistic Regression is a simple and interpretable linear classifier. Decision Trees perform nonlinear splits; however, they are unstable, while Random Forest is based on averaging of the trees, and Gradient Boosting constructs trees one-by-one to fix their mistakes.

The data set was divided into 80% training data (1,176 samples) and 20% test data (294 samples) with the stratified sampling strategy that helps to maintain the proportion of the target class. Random seed was set to 42 for reproducibility of the results. The numeric predictors were median-imputed and normalized; while the categorical predictors were imputed by the mode value and converted into dummy variables. All preprocessing procedures were applied inside the model pipeline to prevent information leakage. Class balancing was performed for Logistic Regression and Decision Tree classifiers, balanced subsamples of classes were used for the Random Forest; the default settings of the Gradient Boosting classifier were used.

3.4 Evaluation Metrics

Accuracy is the ratio of correct classifications but may be misleading in case of class imbalance. Precision calculates how many attrition cases have been correctly predicted while recall calculates how many actual cases have been found. The F1-score takes into account both precision and recall. ROC-AUC is a measure of ranking discrimination and average precision represents the precision-recall curve.

Stratified five-fold cross-validation was also performed on the training data set. Mean F1-score, its standard deviation, mean ROC-AUC, mean recall, and mean precision were noted down. Cross-validation was not done using the test data set that had been held out. Permutation importance was estimated for the model with the highest F1 score by shuffling all raw features individually and calculating the average reduction in test F1-score.

4. Results and Discussion

In this section, the results achieved through the analysis of the IBM HR Analytics Employee Attrition dataset are presented and discussed. Results are analyzed to learn about patterns of employee attrition, analyze the predictive capabilities of supervised machine learning models, and determine the variables that lead to the successful prediction of attrition. At first, descriptive analysis is performed to investigate attrition among overtime employees and job positions. Further, the four classification methods are assessed based on several metrics and cross-validation to analyze the predictive power and reliability of these models. In addition, feature importance is analyzed to determine the variables useful for the prediction of attrition. Lastly, results are interpreted in relation to human resources management and its applications in employee retention.

4.1 Workforce Patterns

Of the 1,470 data entries, 237 belonged to the attrition class while 1,233 belonged to the retained employees class. Consequently, the attrition ratio was 16.12% in this dataset, which means that the positive class comprised a minority of the total data.

Table 1. Dataset and Attrition Summary

Measure	Result
Total employee records	1,470
Employees labelled “left”	237
Employees labelled “stayed”	1,233
Overall attrition rate	16.12%
Predictor features used	30
Training records	1,176
Test records	294

Employee attrition for those who worked overtime was at 30.53%, while the figure was 10.44% for employees who did not work overtime. Clearly, there is a big descriptive difference between these two groups, making overtime a good variable to consider when conducting modeling. However, the existence of such a difference does not indicate a causal relationship between overtime and attrition.

Table 2. Selected Descriptive Findings

Group or measure	Attrition rate / median
Overtime: Yes	30.53%
Overtime: No	10.44%
Sales Representative attrition	39.76% (33 of 83)
Median monthly income, stayed	5,204
Median monthly income, left	3,202

The largest attrition rate in terms of roles was recorded among Sales Representatives with 39.76%. The above results can be used to conduct further analysis regarding the demands of the roles, incentives, schedule, and career path, but the number of people involved in this category is fairly low compared to the whole database. The median salary per month for employees who stayed in the company was 5,204 while that for those who left was 3,202.

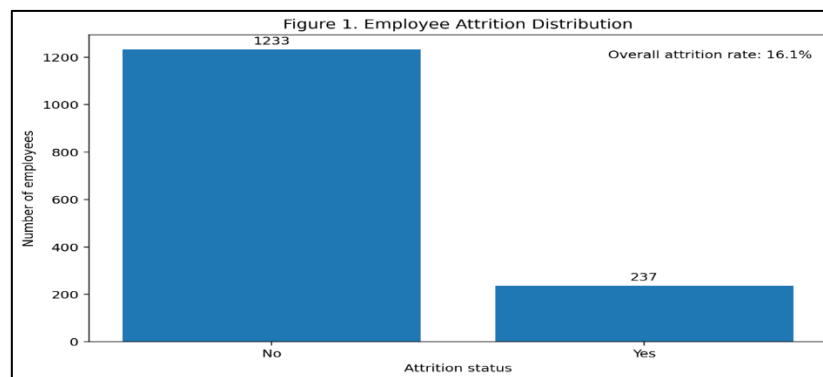


Figure 1 here: Employee Attrition Distribution

The above diagram depicts the distribution of those employees that remained in the firm versus those that left. Out of the total number of 1,470 employees, 1,233 (83.88%) remained, whereas 237 (16.12%) left. Therefore, there is a disparity between the two classes, which must be considered in the assessment of machine learning algorithms.

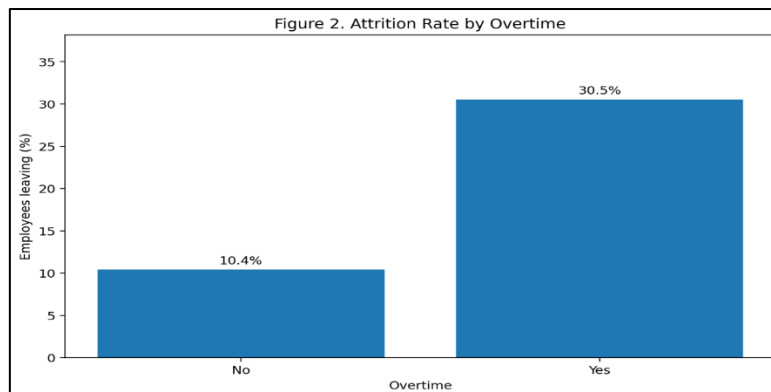


Figure 2. Attrition Rate by Overtime

The graph shows the comparison of employee attrition rate according to the working overtime. It is seen that there was a high attrition rate of 30.53% for those employees who worked overtime while only 10.44% for the other employees who were not working overtime.

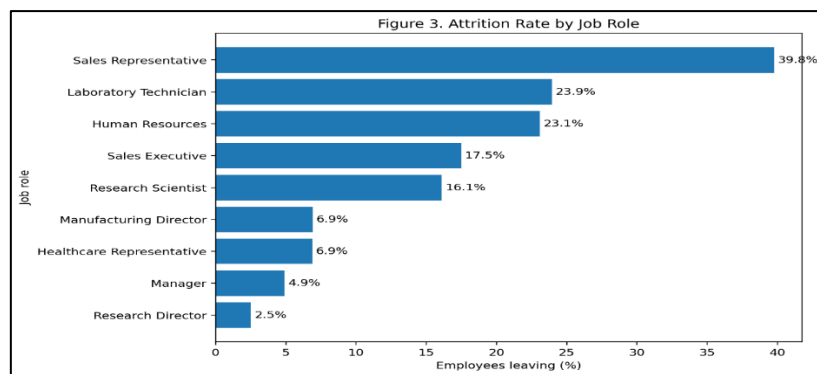


Figure 3. Attrition Rate by Job Role

The above figure displays employee attrition rates for various positions in the organization. The highest attrition rate was found to be among Sales Representatives, with an attrition rate of about 39.76% (33 out of 83 employees). Such a difference indicates that the trends in attrition vary from position to position.

4.2 Comparative Model Performance

Table 3 gives an overview of the test results for the held-out data. Logistic Regression showed the best F1-score (0.451) and recall (0.638). The model's precision was 0.349, which means that a large number of people identified by Logistic Regression left the company. Gradient Boosting showed the highest accuracy (0.861) and almost the best ROC-AUC (0.805), but the recall was only 0.213. Therefore, high accuracy does not equal high detection of attrition. The accuracy of Random Forest was 0.840, and the recall – 0.234. Decision Tree gave the worst ROC-AUC (0.598) and F1-score (0.281).

Table 3. Test-Set Performance of Supervised Learning Models

Model	Accuracy	Balanced accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Logistic Regression	0.752	0.706	0.349	0.638	0.451	0.803	0.561
Decision Tree	0.670	0.562	0.216	0.404	0.281	0.598	0.241
Random Forest	0.840	0.595	0.500	0.234	0.319	0.787	0.439
Gradient Boosting	0.861	0.598	0.714	0.213	0.328	0.805	0.474

The results demonstrate that the choice of model should depend on the HR objective. Recall may be more important for supportive communication, while precision might be more critical for expensive mistakes. HR professionals should pick their thresholds based on the relative cost of mistakes versus true positives. This research uses default thresholds.

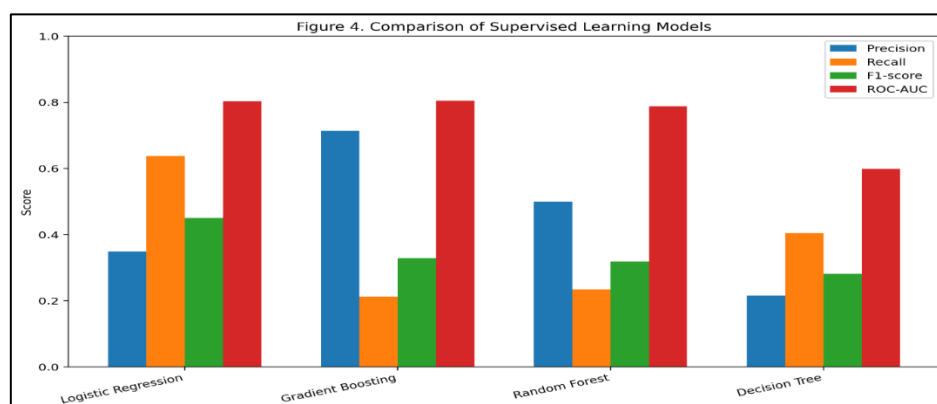


Figure 4. Comparison of Supervised Learning Models

This graph shows the comparison of Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting through various evaluation metrics, including accuracy, precision, recall, F1 score, and ROC-AUC. The Gradient

Boosting model performed the best among all by getting the highest accuracy rate of 86.1%, whereas the Logistic Regression gave the highest recall (63.8%) and F1 score (45.1%).

4.3 Cross-Validation and Feature Importance

From cross-validation performed on the training set, it can be observed that Logistic Regression proved itself to be a good baseline model. Its mean F1-score is 0.492, while its mean ROC-AUC is 0.827, and its standard deviation is 0.034. Gradient Boosting has achieved a mean F1-score of 0.451 and a mean ROC-AUC of 0.813; the same scores for Random Forest are 0.430 and 0.812, correspondingly. Decision Tree managed to produce a mean F1-score of 0.343 and mean ROC-AUC of 0.643. Cross-validation and test scores differ due to different samples used. From the cross-validation results, it may be concluded that Logistic Regression is worth trying, though the choice of model is not confirmed yet.

The Permutation Importance measure for the Logistic Regression model showed an importance ranking of OverTime as the first one with an average reduction of F1 test of 0.1234 upon random permutation. NumCompaniesWorked took the second place at 0.0227, and BusinessTravel took the third at 0.0176. The rest of the features have shown small impacts on the predictive performance. The difference between the feature ranked as the first and the second is quite significant in the fitted model. Still, correlated features may contain the same information, so the Permutation Importance measure may underestimate some features' impact.

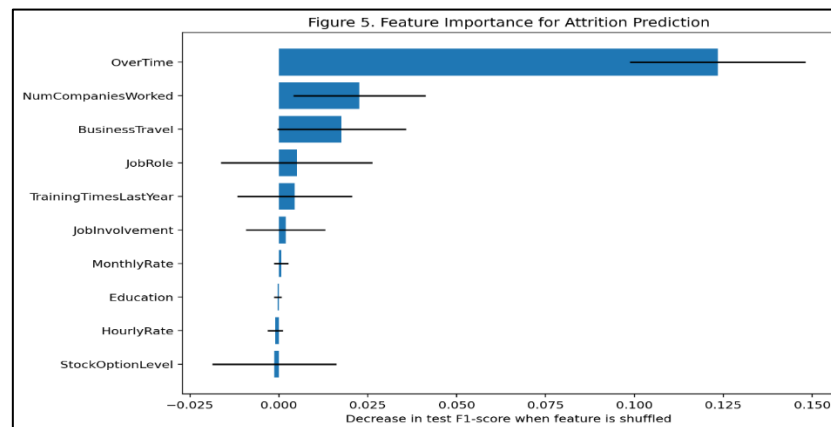


Figure 5. Feature Importance for Attrition Prediction

The chart below is indicative of the importance of various features in the prediction of attrition in employees using the Logistic Regression model. The feature with the greatest importance was found to be OverTime, followed by NumCompaniesWorked and BusinessTravel. This shows that job-related variables can be very helpful in the identification of any pattern related to attrition. However, feature importance only denotes association and not causation.

4.4 Implications for HRM

Some practical applications for workforce analytics from the findings can include:

- The trend of working overtime will call for an analysis of the distribution of workloads, staffing patterns, recovery time and feedback from workers.
- The very high turnover of Sales Reps calls for a targeted diagnostic of the sales representative roles looking at targets, incentives, support and career progression.
- Income level and employment history should be reviewed together with the satisfaction and work conditions factors, and not independently. The aim is to improve the conditions at work and employee retention services and not to judge employees as being disloyal.

Another finding is that the measurement of performance is a managerial decision in addition to being technical. The logistic regression model identified fewer attritions but had a lower precision which means that there were

more false alarms. The HR management teams need to make a decision on whether it is better to provide support earlier or avoid making false alarms. There will be need to validate the models in prospective, monitor for any drift, do subgroup evaluation and enforce privacy.

5. Conclusion and Future Scope

This research has analyzed workforce attrition trends and compared four supervised learning methods on the basis of IBM HR Analytics dataset. The dataset included 1,470 samples, from which 16.12% cases of attrition have been recorded. According to exploratory data analysis, attrition rates were greater for employees with overtime and the highest role-specific attrition rate was observed in Sales Representatives. In the model comparison stage, Logistic Regression demonstrated the highest test F1-score (0.451), recall (0.638) and mean F1-score during cross-validation (0.492). Gradient Boosting had the highest test accuracy (0.861) and slightly better ROC-AUC (0.805). Overtime was the most important feature according to permutation importance approach.

However, there are significant limitations associated with the current study. Firstly, the dataset is generated and rather small, the target variable is imbalanced and the models have not been extensively tuned. Single train-test split may lead to different estimation and feature importance does not imply causal relationships. Thus, results may be regarded only as a baseline experiment.

Future research may evaluate the effect of threshold tuning and cost-sensitive training, compare other machine learning techniques, and conduct repeated or nested cross-validation. Future researchers must explore fairness, examine interpretability with SHAP, and validate the models using suitably governed longitudinal employment data. Human Resource Management departments could also experiment with the effectiveness of these interventions in improving the employee experience and retention rate.

References

1. M. Al Akasheh, E. F. Malik, O. Hujran, and N. Zaki, "A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review," *Expert Systems with Applications*, vol. 238, article 121794, 2024. <https://doi.org/10.1016/j.eswa.2023.121794>
2. "Understanding the adoption and institutionalization of workforce analytics: A systematic literature review and research agenda," *Human Resource Management Review*, vol. 33, no. 4, article 100985, 2023. <https://doi.org/10.1016/j.hrmr.2023.100985>
3. F. K. Alsheref, I. E. Fattoh, and W. Mead, "Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms," *Computational Intelligence and Neuroscience*, 2022, article 7728668. <https://doi.org/10.1155/2022/7728668>
4. "Machine Learning in HR Analytics: A Comparative Study on the Predictive Accuracy of Attrition Models," in *2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*, 2024. <https://doi.org/10.1109/DICCT61038.2024.10533064>
5. H. Talebi, A. Khatibi Bardsiri, and V. Khatibi Bardsiri, "Machine Learning Approaches for Predicting Employee Turnover: A Systematic Review," *Engineering Reports*, 2025. <https://doi.org/10.1002/eng2.70298>
6. "Machine learning with real-world HR data: mitigating the trade-off between predictive performance and transparency," *The International Journal of Human Resource Management*, 2024. <https://doi.org/10.1080/09585192.2024.2335515>
7. "A cluster-based human resources analytics for predicting employee turnover using optimized Artificial Neural Networks and data augmentation," *Decision Analytics Journal*, vol. 11, article 100461, 2024. <https://doi.org/10.1016/j.dajour.2024.100461>
8. P. Kumar, S. B. Gaikwad, S. T. Ramya, T. Tiwari, M. Tiwari, and B. Kumar, "Predicting Employee Turnover: A Systematic Machine Learning Approach for Resource Conservation and Workforce Stability," *Engineering Proceedings*, vol. 59, no. 1, article 117, 2023. <https://doi.org/10.3390/engproc2023059117>