# Predictive Analytics for Workforce Stability: Identifying Attrition Factors with Supervised Machine Learning

Michelle Zheng
Master of Science in Data Science
Regis University, Denver, CO, USA
[mzheng001@regis.edu]

**Abstract**

High employee turnover poses significant financial and organizational challenges across industries. Using the IBM Watson HR Employee Attrition and Performance synthetic dataset of 1,470 employee records with 35 features, the project encompasses data acquisition, preprocessing, exploratory data analysis (EDA), model development, and evaluation using Python. Two algorithms were compared: Logistic Regression and Random Forest. The results show that Logistic Regression achieved a higher Area Under the Curve (AUC) of 0.8374 on the test set, demonstrating stronger discriminative ability for identifying employees at risk of leaving, while Random Forest achieved higher raw accuracy at 84.69%. Key drivers of attrition identified include monthly income, age, overtime status, total working years, and years at the company. The results aim to support proactive, data-driven retention strategies while acknowledging limitations inherent in synthetic historical data.

## I. Introduction

Employee attrition is the gradual reduction of a workforce as employees leave an organization and are not immediately replaced. It can be voluntary, like when an employee resigns, or involuntary, like layoffs or terminations. No matter the cause, attrition creates real costs for organizations, including recruiting and onboarding new employees, lost of institutional knowledge and decreased productivity during reduced headcount periods. Research shows that replacing an employee can cost between 1.5 and 2 times that employee's annual salary, which can amount to millions of dollars for large organizations (Manafi Varkiani et al., 2025) [1]. Beyond financial costs, attrition affects whole industries and economies. When skilled workers leave at high rates, companies lose competitive advantages and entire sectors can experience talent shortages. For example, in healthcare, high nurse turnover can directly affect patient outcomes. In technology, losing experienced engineers slows product development. Understanding why employees leave and predicting who is likely to leave before they do is a valuable problem that data science can help solve.

Having personally witnessed high employee turnover, the interest behind this topic comes from seeing first-hand how reactive retention strategies often fail. Most organizations only respond to attrition after it happens, rather than identifying warning signs early. This project investigates factors of employees at risk of leaving by building machine learning models that can flag employees with elevated risks, giving organizations the opportunity to intervene before the departure occurs.

## II. Problem Statement

The problem statement is centered around how machine learning models be effectively developed and applied to predict employee attrition and identify the multi-variable drivers that lead employees to leave?

## III. Related Work

Govindarajan et al. (2025) [2] conducted a comparative analysis of machine learning models for predicting attrition using the same IBM HR Analytics dataset used in this project. Their study evaluated several algorithms like Decision Trees and found that ensemble methods generally outperformed simpler classifiers in terms of accuracy. Their work focused primarily on overall accuracy as the main performance metric, which can be misleading when dealing with datasets of binary classifications that very skewed. The limitation can be addressed by also evaluating AUC, precision, recall and F1-score to provide a more complete performance picture.

Manafi Varkiani et al. (2025) [1] investigated attrition prediction with a focus on identifying and explaining the key determinants of turnover. Their findings highlighted the importance of interpretability in HR analytics, meaning that it is not enough to predict who will leave. Companies also need to understand why and this insight directly motivates the feature importance analysis performed in this project using the Random Forest model.

Hoffman and Tadelis (2021) [3] examined the role of people management quality in employee attrition, finding that poor managerial performance was significantly associated with higher turnover rates among subordinates. Their work emphasizes that attrition is not driven solely by compensation and that the quality of the manager-employee relationship is a powerful independent factor. This is supported with the inclusion of variables like years with current manager and relationship satisfaction in this project.

## IV. Data Description

### A. Dataset Overview

The dataset used in this project is the IBM Watson HR Employee Attrition and Performance dataset. A synthetic dataset that contains 1,470 employee records and 35 features covering a large range of employee attributes, including demographic variables like age, gender, and marital status; compensation variables like monthly income; job characteristics like job role and department; satisfaction measures like job satisfaction; and career history variables like total working years and years with current manager.

The target variable is “Attrition,” a binary indicator of whether an employee is a past employee company (Yes) or current employee (No). Of the 1,470 records, 1,233 employees (83.88%) were still with the company at the time of data collection, while 237 employees (16.12%) had left.

### B. Data Ethics and Governance

Real employee data is difficult to obtain for research purposes. Companies do not make personnel records public and accessing non-synthetic demographic data typically requires non-disclosure agreements, making it inaccessible for open academic research. The use of a synthetic dataset addresses these concerns while still preserving the statistical properties needed for analysis (Alqahtani et al., 2025) [4].

The use of synthetic data itself raises ethical considerations worth acknowledging. Susser et al. (2024) [5] note that synthetic health and organizational data carries both promise and peril. While it removes privacy risks associated with real identifiable individuals, it can introduce its own biases depending on how the synthetic data was generated. Machine learning models trained on synthetic data may not generalize perfectly to real organizational environments. Cooper and Coetzee (2020) [6] further emphasize that even with publicly available data, researchers must consider purpose, potential harm and transparency in how results are communicated.

An example of re-identification risk from real data is the Netflix Prize dataset challenge, in which researchers were able to reverse-engineer anonymized customer records to identify individuals (Narayanan & Shmatikov, 2006) [7]. This case reinforces why synthetic data is the responsible choice for this type of research when real employee records are not available through proper consent channels.

## V. Preparation and Methodology

### A. Tools and Libraries

The project was implemented in Python. Core libraries used include `pandas` and `NumPy` for data manipulation, `Matplotlib` and `Seaborn` for static data visualization, `Plotly Express` for interactive visualization, `SciPy` for statistical testing, and `scikit-learn` for machine learning modeling, preprocessing, and evaluation.

### B. Exploratory Data Analysis

Sales showed the highest attrition rate relative to its size, followed by Human Resources, while Research and Development had comparatively lower turnover. Figure 1 illustrates the distribution of current and past employees across departments.

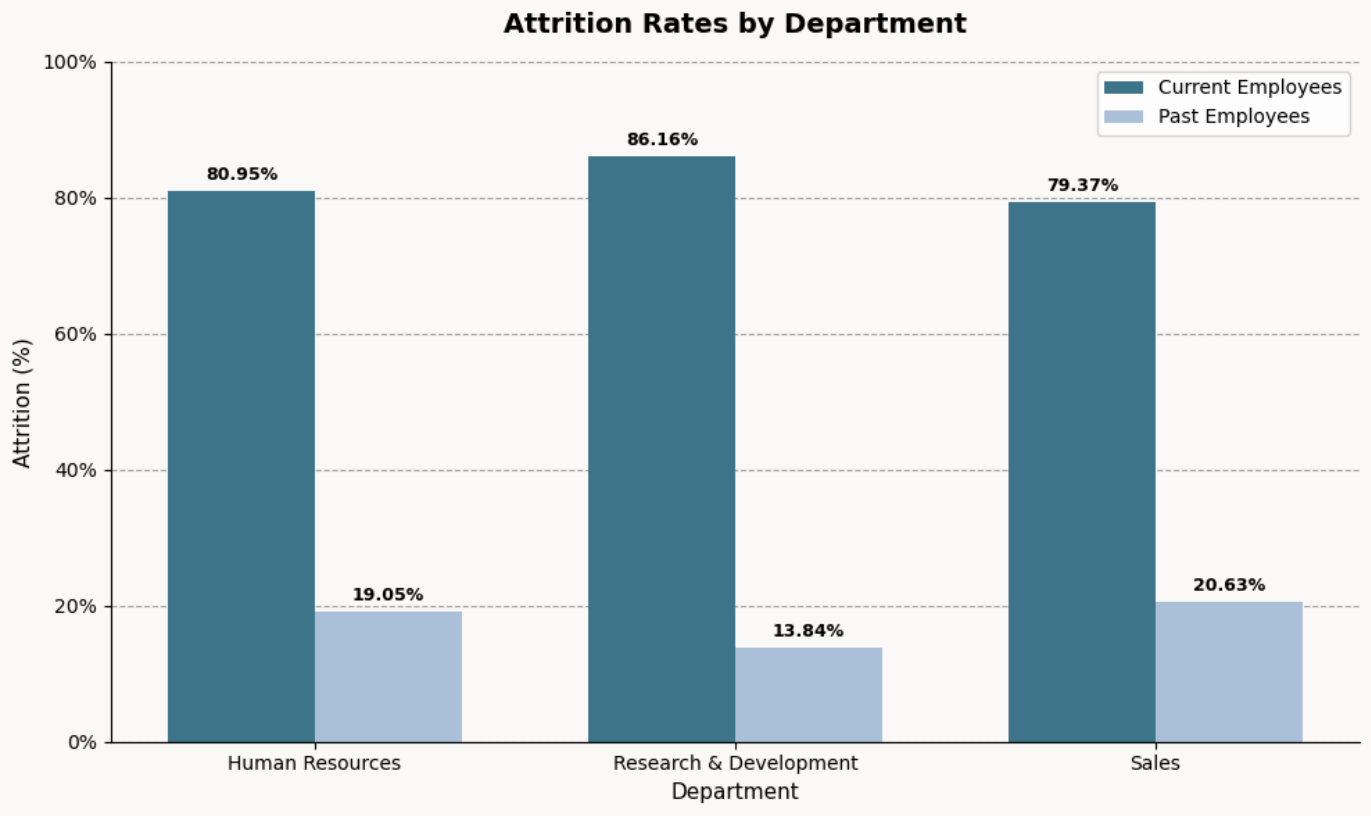


Fig. 1. Attrition rates by department.

Younger employees in their mid-to-late twenties and early thirties accounted for the largest share of departures, with attrition dropping off noticeably after age 35. Younger workers might be more likely to change positions before settling into long-term roles (Hoffman & Tadelis, 2021) [3]. Figure 2 shows attrition counts across the age spectrum.

Employees who left were most concentrated among those with 0 to 2 years with their current manager, suggesting that newer or recently reassigned manager-employee relationships are more fragile. Figure 3 shows attrition counts by years worked with the current manager.

### C. Correlation Matrix

The correlation matrix highlighted a few key relationships. Attrition correlated negatively with job level, monthly income, total working years and stock option level, suggesting that more senior, higher-compensated employees are less likely to leave. Overtime showed a positive correlation with attrition, reinforcing its role as a risk factor. Figure 4 displays the full correlation matrix.

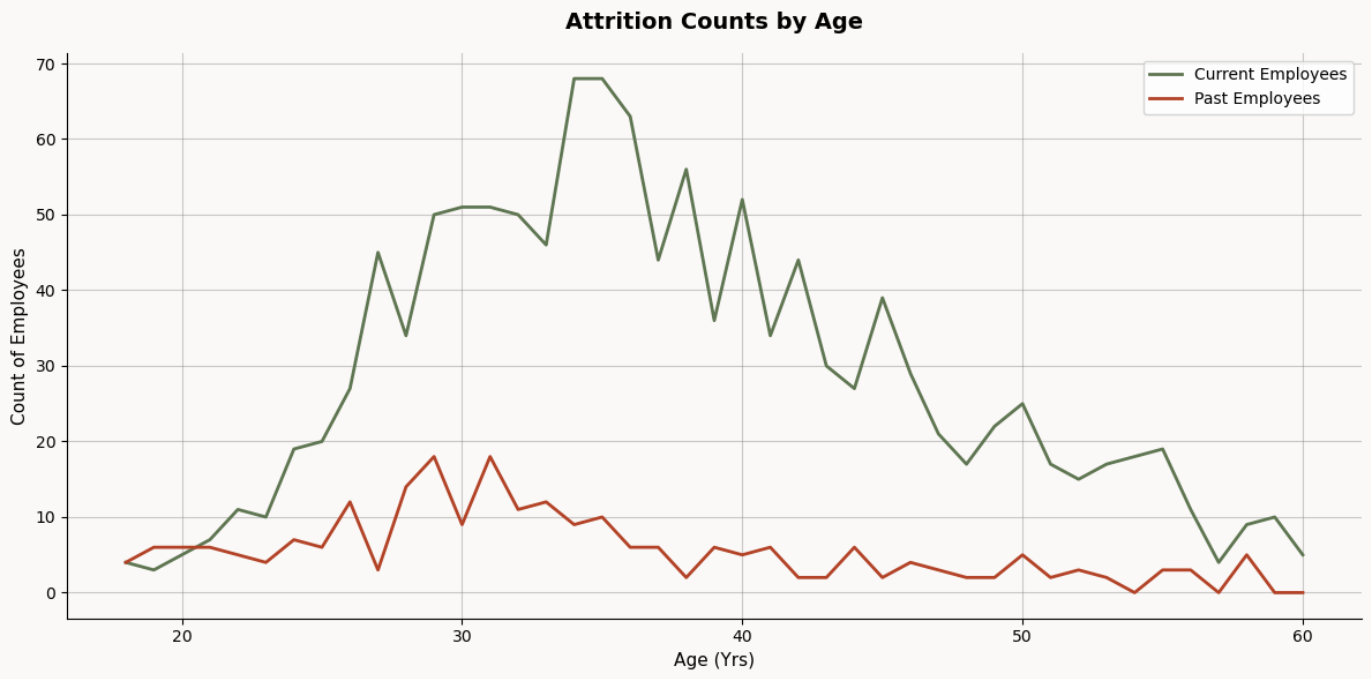


Fig. 2. Attrition counts by employee age.

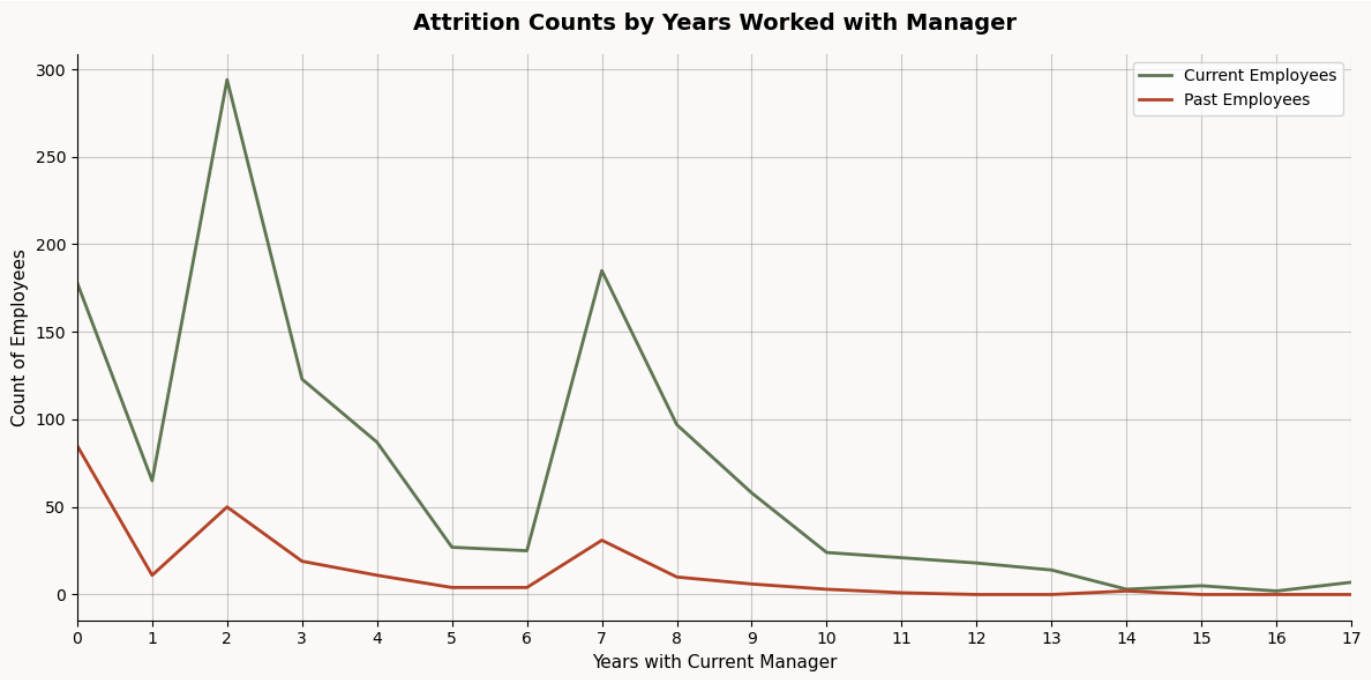


Fig. 3. Attrition counts by years with current manager.

## *D. Steps before Modeling*

Before modeling, the dataset went through some preprocessing steps. Four columns were removed due to providing no analytical value: `EmployeeCount`, `StandardHours`, `Over18`, and `EmployeeNumber`. Variables representing ordered categorical ratings, like job satisfaction, environment satisfaction, job level and performance rating, where cast as categorical types.

A skewness analysis was then conducted on all remaining numerical features. Nine columns with an

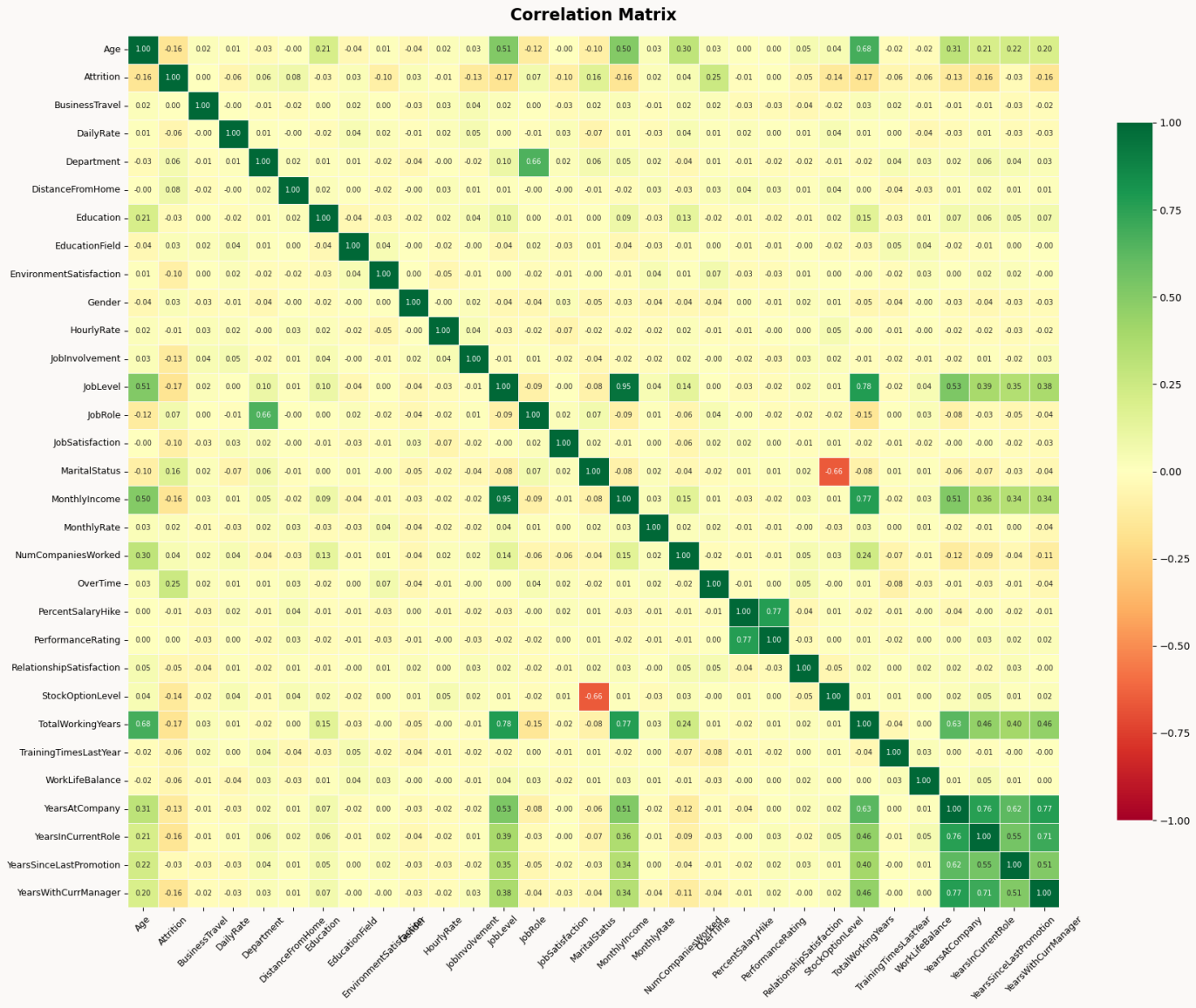


Fig. 4. Correlation matrix across all numeric and encoded variables.

absolute skewness value greater than 0.75 were log-transformed to compresses extreme values and brings distributions closer to normality.

### *E. Statistical Testing*

Also prior to modeling, an independent samples *t*-test was conducted to check whether the difference in average monthly income between employees who left and those who stayed was statistically significant. The *t*-test returned a test statistic of $t = -7.7477$ with a $p$-value effectively equal to zero ($p < 0.0001$), rejecting the null hypothesis. Past employees earned significantly lower average monthly incomes than current employees, confirming that income is a meaningful predictor of attrition.

### *F. Model Training and Evaluation*

The dataset was split into training (60%), validation (20%), and test (20%) sets using stratified sampling to preserve the class distribution for the sets. All categorical features were one-hot encoded using `pandas get_dummies` with `drop_first=True` to avoid multicollinearity.

Two classification models were trained and evaluated. These two models were chosen because logistic regression tends to achieve higher overall accuracy when data variance increases, while random forest yields a higher true positive rate, which makes them complementary options where identifying employees likely to leave is as important as overall accuracy (Kirasich et al., 2018) [8]. Logistic Regression was implemented with a balanced class weight setting to counteract the class imbalance, using the LBFGS solver with a maximum of 1,000 iterations. Random Forest was tuned using `RandomizedSearchCV` with 50 iterations and 5-fold cross-validation, choosing AUC as the scoring metric. The best hyperparameters found were 336 estimators and a minimum of 17 samples to split an internal node. Models were evaluated and compared using accuracy, AUC, precision, recall, F1-score and average precision.

## VI. Model Results and Comparison

### A. Logistic Regression

The Logistic Regression model achieved an accuracy of 78.91% on the validation set and 79.25% on the test set. It achieved AUC scores of 0.8413 (validation) and 0.8374 (test), which shows a strong discriminative ability between employees who left and those who stayed. The recall for the past employees was 0.6875 on the validation set and 0.7660 on the test set, meaning the model correctly identified approximately 69–77% of employees who had actually left, an important metric for a retention-focused application where missing a high-risk employee could come at a high cost.

### B. Random Forest

The Random Forest model achieved higher raw accuracy than Logistic Regression: 83.33% on the validation set and 84.69% on the test set. Although this accuracy came at a significant cost to recall for the past employees classification. On the validation set, the Random Forest correctly identified only 2.08% of employees who had left (recall $= 0.0208$), and on the test set, only 4.26% (recall $= 0.0426$). The model almost entirely defaulted to predicting "current employee" for most records, artificially inflating its accuracy while providing little practical value for identifying who is at risk of leaving. AUC scores of 0.7952 (validation) and 0.7719 (test) were also lower than those of Logistic Regression.

### C. Model Comparison Summary

Table I summarizes all evaluation metrics for both models across the validation and test sets.

TABLE I
Model Comparison — Metric Summary

| Model | Acc. | AUC | Prec. | Recall | F1 | Avg P |
|---|---|---|---|---|---|---|
| RF — Val | 83.33% | 0.7952 | 0.3333 | 0.0208 | 0.0392 | 0.4285 |
| RF — Test | 84.69% | 0.7719 | 1.0000 | 0.0426 | 0.0816 | 0.5128 |
| LR — Val | 78.91% | 0.8413 | 0.4125 | 0.6875 | 0.5156 | 0.5824 |
| LR — Test | 79.25% | 0.8374 | 0.4186 | 0.7660 | 0.5414 | 0.6587 |

The precision-recall curves further shows the difference between the two models. Logistic Regression maintained a stronger balance between precision and recall across different classification thresholds. While Random Forest showed better precision at very high-confidence predictions but collapsed rapidly in recall. Figure 5 shows the precision-recall curves for both models.

### D. Feature Importance

The Random Forest model provides feature importance scores indicating which variables contributed most to its predictions. Figure 6 shows the top 15 most important features.

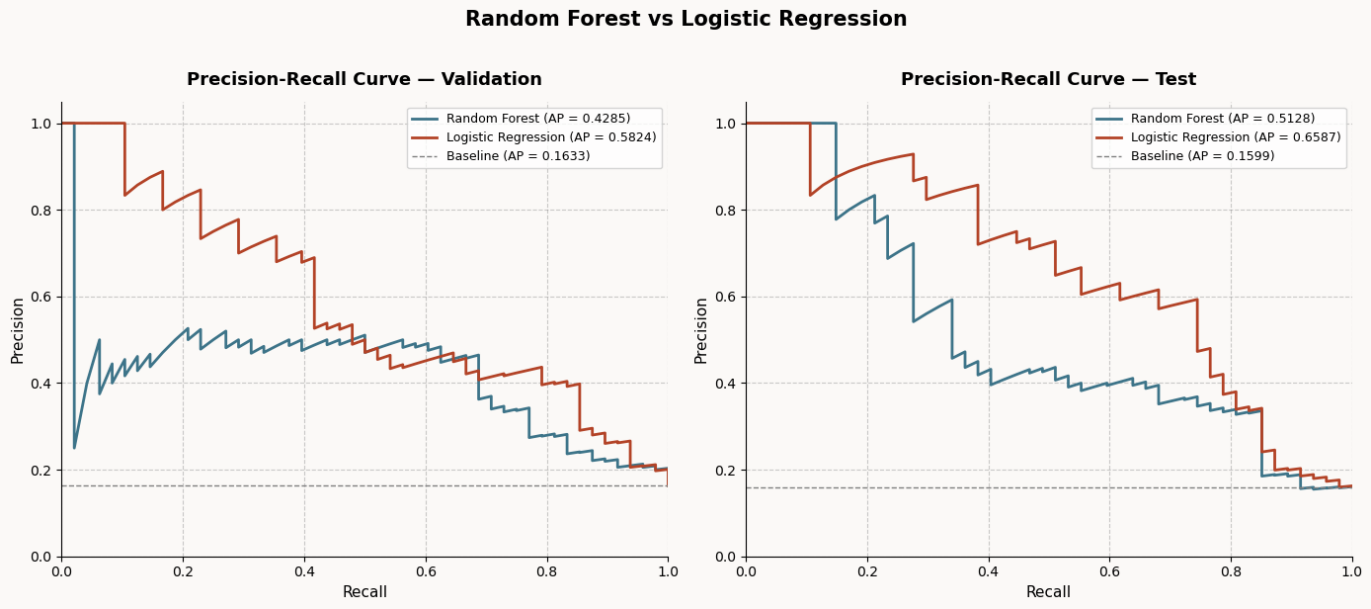


Fig. 5. Precision-recall curves: Random Forest vs. Logistic Regression on validation and test sets.

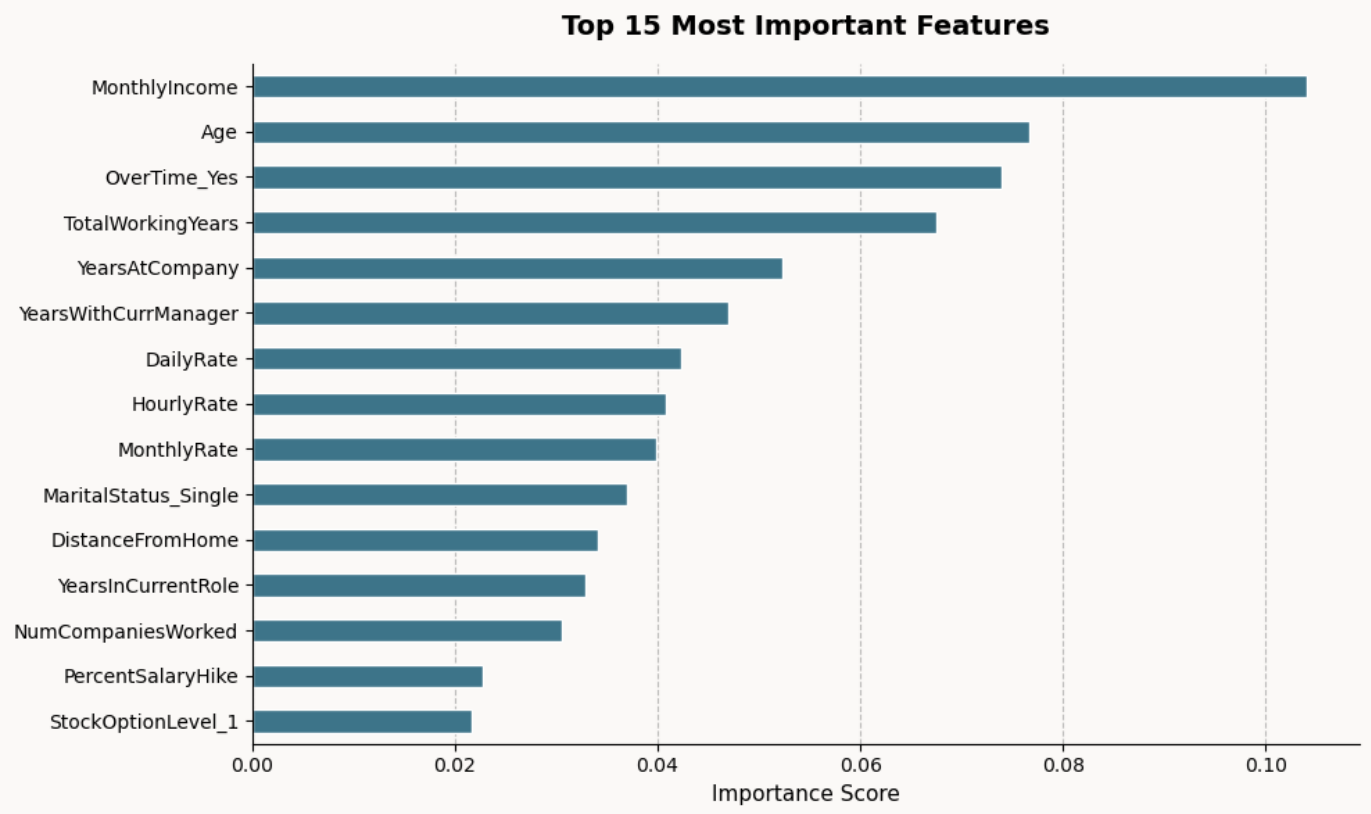


Fig. 6. Top 15 most important features from the Random Forest model.

## VII. Key Findings

Compensation is a statistically confirmed driver of attrition. The *t*-test showed that employees who left earned significantly lower monthly incomes than those who stayed, and `MonthlyIncome` was the single most important feature in the Random Forest model.

Overtime is a major risk factor. Employees required to work overtime were more likely to leave, suggesting issues with workload management does not respect work-life balance. Younger employees and those early in their tenure at the company, or with their manager showed higher attrition rates, pointing to the value of proactive manager support in the first two years.

Model selection must match the problem. Logistic Regression correctly identified 77% of at-risk employees on the test set, compared to just 4% for Random Forest, making it a superior choice for this use case. For imbalanced data, recall and AUC matter more than raw accuracy, a finding consistent with Kirasich et al. (2018) [8].

## VIII. Conclusion

The Logistic Regression model emerged as the more effective choice for this classification problem, achieving an AUC of 0.8374 and a recall of 76.60% for the attrition class on the test set. These results indicate that the model successfully identified approximately three out of every four employees who had left. Factors like monthly income, overtime, age, total working years, and years with the current manager were identified as the most influential predictors of attrition, pointing to specific areas where organizations can take targeted action.

Limitations should be acknowledged, where the IBM dataset is synthetic, meaning that models trained on it may not directly transfer to real organizational data without retraining.

Future work could address these limitations. Integrating real employee data through anonymized, consent-based access with organizations would strengthen model generalizability. Incorporating natural language processing on exit interview data or engagement survey responses, an approach explored by Lee and Algarra (2025) [9] could add interpretive depth beyond structured HR variables. Developing a real-time scoring system that flags individual employees for HR review, could transform this analytical framework into a deployable model for workforce stability management.

## References


[1] S. Manafi Varkiani, F. Pattarin, T. Fabbri, and G. Fantoni, "Predicting employee attrition and explaining its determinants," *Expert Systems with Applications*, vol. 272, p. 126575, 2025.

[2] R. Govindarajan, N. K. Kumar, S. R. P, S. P. E, D. B, and P. K. G, "Predicting employee attrition: A comparative analysis of machine learning models using the IBM human resource analytics dataset," *Procedia Computer Science*, vol. 258, pp. 4084–4093, 2025.

[3] M. Hoffman and S. Tadelis, "People management skills, employee attrition, and manager rewards: An empirical analysis," *Journal of Political Economy*, vol. 129, no. 1, pp. 243–285, 2021.

[4] H. Alqahtani, H. Almagrabi, and A. Alharbi, "Dataset for predictive modelling and analysis of employee attrition and retention," *Data in Brief*, vol. 63, p. 112242, 2025.

[5] D. Susser, D. S. Schiff, S. Gerke, L. Y. Cabrera, I. G. Cohen, M. Doerr, J. Harrod, K. Kostick-Quenet, J. McNealy, M. N. Meyer, W. N. Price, and J. K. Wagner, "Synthetic health data: Real ethical promise and peril," *Hastings Center Report*, vol. 54, no. 5, pp. 8–13, 2024.

[6] A. K. Cooper and S. Coetzee, "On the ethics of using publicly-available data," in *Lecture Notes in Computer Science*, 2020, pp. 159–171.

[7] A. Narayanan and V. Shmatikov, "How to break anonymity of the Netflix prize dataset," 2006.

[8] K. Kirasich, T. Smith, and B. Sadler, "Random forest vs logistic regression: Binary classification for heterogeneous datasets," *SMU Data Science Review*, vol. 1, no. 3, 2018. [Online]. Available: https://scholar.smu.edu/datasciencereview/vol1/iss3/9

[9] F. Lee and A. Algarra, "Leveraging topic modeling to predict and prevent employee attrition," *Information Systems Education Journal*, vol. 23, no. 4, pp. 59–84, 2025.

[10] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, no. 1, 2018.

[11] J. Kim, "The effect of mismanagement of poor performers on their coworkers' turnover intentions," *Public Personnel Management*, vol. 55, no. 1, pp. 118–144, 2025.

[12] G. N. Srivastava, H. Sharma, R. N. Agarwal, and A. K. Jain, "Analyzing employee attrition of research and development firms using mixed methods," *Cogent Business & Management*, vol. 12, no. 1, 2025.