Predicting employee departures using machine learning techniques: HR data analysis

Abdulrahman Radan, Mosab Al-Hobaishi, Malik Al-Masbahi, Mahmood Al-Yamani, Abdulgoni Al- Gholi, Al-Hamza Antar, Jarman Jarman

Article:

info Article history: Received 2 February 2025 Revised

Keywords:
DNLC approach Deep Neural Network learning Cancer prediction Cancer classification Feature selection
, Predicting, employee departures, deep learning, data science, data analysis, dataset analysis,

Abstract:
Employee attrition poses a significant challenge for organizations by driving up costs, depleting talent, and reducing productivity. This study employs machine learning techniques to predict employee attrition using the IBM HR Analytics Employee Attrition & Performance dataset. The dataset comprises diverse features, including demographics, job satisfaction levels, work environment conditions, and performance metrics. Logistic regression and random forest models are applied to identify the critical factors influencing attrition and develop an accurate predictive model. Results reveal that random forest outperforms logistic regression in terms of prediction accuracy. This research provides actionable insights for human resource management, enabling organizations to proactively address employee turnover and implement targeted strategies to improve retention. By showcasing the value of data-driven methods, this study offers a foundation for further exploration of predictive analytics in workforce management.

3. To propose actionable recommendations for human resource teams to strengthen employee retention.

This paper is organized as follows: Section 2 provides an overview of related research in employee attrition prediction. Section 3 describes the dataset and preprocessing techniques. Section 4 explains the methodology, including the machine learning models and evaluation metrics employed. Section 5 presents and discusses the results. Finally, Section 6 concludes the study and suggests avenues for future research.
In conclusion, this introduction frames the study as a comprehensive exploration of employee attrition prediction, emphasizing the significance of data-driven approaches. By leveraging advanced machine learning techniques, this research contributes meaningful insights to human resource management, aiding organizations in reducing attrition and fostering a more stable and productive workforce.

---

i. . Related works

Predicting Employee Turnover
Study Results:
1. Model Accuracy:
• Several machine learning algorithms were used, and each performed differently in terms of accuracy and efficiency.
• Best performing models:
o Random Forest: It was the most accurate in prediction due to its ability to handle complex and multidimensional data.
o Gradient Boosting: It provided excellent results in discovering subtle patterns in the data.
o Logistic Regression: It performed well in explaining the relationship between different factors and the likelihood of leaving, but it was less accurate than other models.
2. Main Influencing Factors:
The factors that most influenced employees' decision to leave were identified as:
• Job satisfaction level: It was the most influential factor, as the data showed that low satisfaction significantly increased the likelihood of an employee leaving.
• Work-life balance: Long hours or overtime had a negative impact on employee satisfaction.
• Promotions and future opportunities: The lack of opportunities for promotions or career

Introduction:

Employee attrition, defined as the voluntary or involuntary departure of employees from an organization, remains a pressing issue for businesses worldwide. High rates of attrition often result in considerable financial costs, decreased productivity, and diminished employee morale. Industry estimates suggest that replacing an employee can cost between 50% and 200% of their annual salary, emphasizing the need for organizations to predict and address attrition through effective strategies. This research aims to utilize data science techniques to better understand and predict employee attrition, offering actionable insights that support workforce stability and mitigate turnover.
The complexity of employee attrition arises from its dependence on multiple factors, including job satisfaction, compensation, work environment, and opportunities for professional growth. Conventional methods for addressing attrition tend to be reactive and less effective compared to proactive, data-driven approaches. Through the application of machine learning models, this study seeks to uncover patterns and predictors of attrition, enabling organizations to implement preemptive measures that enhance employee retention.
This research stands out for its potential to revolutionize human resource management by integrating predictive analytics into strategic decision-making processes. Using the IBM HR Analytics Employee Attrition & Performance dataset, which provides comprehensive details on employee demographics, job roles, and performance indicators, the study aims to construct accurate and interpretable predictive models. Logistic regression and random forest methodologies are employed, chosen for their robustness in handling complex datasets and their ability to yield actionable insights.

The objectives of this study are threefold:

---

1. To develop a reliable predictive model for employee attrition.
2. To identify the key factors contributing to turnover.

• Practical applications: Improving human resources strategy, designing job loyalty programs, and providing a stimulating work environment.

---

3. Factors affecting employee retention:
The paper focuses on a number of factors that can be measured using predictive analytics:
• Job satisfaction level: It is affected by factors such as wages, work-life balance, and company culture.
• Career advancement: Lack of growth and promotion opportunities increases the likelihood of leaving.
• Periodic evaluation: Employee performance may be an early indicator of the level of engagement with the company.
• Institutional belonging: Employees feel appreciated and respected within the team.

4. Analysis tools used:
The study includes the use of a set of tools and techniques to analyze the data:
• Machine learning: to identify recurring patterns and link them to the likelihood of leaving.
• Linear regression models: to determine the relationship between different factors and the rate of leaving.
• Data visualization techniques: to facilitate the presentation of results to the HR team.

5. Key findings:
• Predictive accuracy: Predictive analytics can achieve an accuracy of more than 80% in predicting employees most likely to leave.
• Preventive analysis: Companies can intervene proactively, such as increasing salaries or providing training opportunities, which reduces turnover rates by up to 30%.
• Most important factors: Job satisfaction, promotion opportunities, and work-life balance were the most influential factors.

6. Practical recommendations:
1. Invest in predictive analytics: Companies should adopt modern systems to analyze employee data.
2. Focus on company culture: Provide a supportive and motivating environment that helps reduce employee turnover.
3. Proactive: Implement proactive plans based on predictions, such as providing incentives or reviewing work policies.
4. Personalize programs: Design personalized retention plans based on the needs of each employee.

advancement within the company was an important factor in making the decision to leave.
• Wages: Employees who felt unfair in salaries were more likely to leave.
• Age and years of work: Younger employees and those who spent less time with the company were more likely to leave.
3. Prediction performance:
• The models showed good ability to predict which employees are most likely to leave, allowing companies to intervene early.
• The weighted distribution in the prediction helped reduce errors associated with the category of employees who actually leave.
4. Additional insights:
• The factor most positively associated with employee retention was appreciation and support from management.
• The relationship with colleagues played a role in an employee's decision to stay or leave, although its impact was less than other factors.
Recommendations based on the results:
• Improve the work environment: Enhance work-life balance and increase satisfaction levels.
• Re-evaluate salaries: To ensure fairness and competition in the market.
• Focus on professional development opportunities: Provide ongoing promotion and training opportunities.
• Conduct periodic surveys: To learn about employees' expectations and needs before they reach the stage of thinking about leaving.
Conclusion about the results:
The results demonstrated that applying machine learning models can help companies identify employees at risk of leaving. If strategic decisions are made based on these findings, companies can significantly reduce turnover, thereby reducing costs and increasing productivity.
Predictive Analytics for Employee Retention:
1. The main problem:
Employee turnover is one of the biggest challenges facing modern organizations, as it leads to material and moral losses, such as high recruitment costs, loss of institutional knowledge, and low morale within the team.
The study focuses on the role of predictive analytics in helping organizations understand the causes of employee turnover, anticipate it, and take proactive measures to reduce it.

---

2. The importance of predictive analytics:
Predictive analytics relies on the use of big data and artificial intelligence techniques to identify patterns and trends leading to employee departure.
• Main benefit: Enables organizations to predict employees most likely to leave and design customized plans to retain them.

- **Removing Outliers:** Using the Z-score method or IQR-based filtering to remove noisy data points that could distort the predictions.
- **Scaling the Data:** Normalization or standardization was applied where necessary to improve the model's convergence during training.

2. Feature Engineering

Feature engineering was performed to enhance model performance and interpretability. The key steps included:

- **Feature Encoding:** Categorical variables were encoded using frequency encoding.
- **Data Balancing:** Since employee attrition data is often imbalanced, balancing techniques were used to improve predictive performance.
- **Feature Selection:** Identifying the most important variables affecting attrition rates.

Feature Selection Techniques

Several feature selection methods were applied:

1. **Chi-Square Test**: Used for categorical variables to determine statistical significance in attrition prediction.
    - Formula: X2=∑(O−E)2EX^2 = \sum \frac{(O - E)^2}{E}
    - Where OO is the observed frequency, and EE is the expected frequency.
2. **Mutual Information**: Measures dependency between variables.
    - Formula: I(X,Y)=∑p(x,y)log⁡p(x,y)p(x)p(y)I(X, Y) = \sum p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
3. **Principal Component Analysis (PCA)**: Reduces dimensionality while retaining variance.
    - Formula: X′=XWX' = X W
    - Where XX is the original dataset and WW is the weight matrix of principal components.

3. Data Balancing Techniques

Since employee attrition datasets are often imbalanced, different resampling methods were explored to ensure model performance remains unbiased.

SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE generates synthetic samples to balance class distribution.

- Formula: xnew=xi+λ(xj−xi)x_{new} = x_{i} + \lambda (x_{j} - x_{i})
- Where xix_{i} and xjx_{j} are two close instances, and λ\lambda is a random number between 0 and 1.

Alternative Methods to SMOTE

1. **Random Oversampling:** Increases minority class samples by duplicating instances.

II. Literature Review

In this study, Data Science and Deep Learning were employed to analyze employee turnover rates and predict the reasons behind employee departures from companies. The research focused on analyzing data to determine whether an employee's resignation was due to company decisions, the nature of the job, or various constraints. A set of models and algorithms were applied to analyze these factors, and a Data Processing and Visualization App was used for data preparation and analysis. Below is a review of previous related studies and techniques used in our research:

Predictive Analysis of Employee Turnover Using Data Science:

Researchers team applied a model based on Machine Learning and Deep Learning algorithms to predict employee turnover based on various characteristics such as length of service, work environment, salaries, and promotions. The researchers utilized Feature Engineering and Data Processing techniques to enhance the accuracy of predictive models.

Using Data Balancing and Feature Analysis to Identify Factors Influencing Employee Turnover:

In a study conducted by team, techniques such as Handling Missing Data, Data Cleaning, and Data Balancing were used to ensure the model accurately reflected the factors influencing employee resignations.

Employee Turnover Analysis Using Deep Learning:

Researchers ream, relied on Deep Neural Networks to analyze HR datasets, where the model demonstrated higher accuracy compared to traditional algorithms. The study emphasized techniques such as Frequency Encoding to select the most influential features affecting employee turnover decisions.

Data Science Approach for Employee Attrition Prediction

1. Data Collection and Processing

To build an effective predictive model for employee attrition, data collection and preprocessing are crucial steps. This study utilizes the IBM HR Analytics Employee Attrition & Performance dataset, which consists of various features such as demographic details, job satisfaction, work environment, and performance metrics. The following preprocessing techniques were applied:

- **Checking for Missing Values:** Identifying and handling missing data to ensure model robustness.
- **Handling Missing Values:** Filling missing values with the mean, median, or using predictive imputation methods.

and performance-related attributes. With a total of 1,471 records, this dataset provides a comprehensive view of factors potentially contributing to employee attrition.

Dataset Features

The following are key features present in the dataset:

- **Age**: Numeric (Discrete) — Represents the age of the employee.
- **Attrition**: Categorical — Indicates whether the employee has left the organization (Yes/No).
- **Business Travel**: Categorical — Frequency of business travel (Travel Rarely, Travel Frequently, No Travel).
- **Department**: Categorical — The department where the employee works (Sales, Research & Development, etc.).
- **Distance From Home**: Numeric (Discrete) — Distance between the employee's home and workplace.
- **Education**: Categorical — Employee's level of education.
- **Education Field**: Categorical — Field of education (Life Sciences, Medical, etc.).
- **Gender**: Categorical — Gender of the employee (Male/Female).
- **Job Level**: Categorical — Represents the seniority level of the job.
- **Job Role**: Categorical — Role of the employee in the organization (Research Scientist, Sales Executive, etc.).
- **Job Satisfaction**: Categorical — Level of job satisfaction.
- **Marital Status**: Categorical — Marital status of the employee (Single, Married, Divorced).
- **Monthly Income**: Numeric (Discrete) — Employee's monthly income.
- **Num Companies Worked**: Numeric (Discrete) — Number of companies the employee has previously worked for.
- **OverTime**: Categorical — Whether the employee works overtime (Yes/No).
- **Percent Salary Hike**: Numeric (Discrete) — Percentage increase in the salary.
- **Performance Rating**: Categorical — Performance rating of the employee.
- **Relationship Satisfaction**: Categorical — Level of satisfaction in workplace relationships.
- **Total Working Years**: Numeric (Discrete) — Total years of professional experience.
- **Training Times Last Year**: Numeric (Discrete) — Number of training sessions attended in the previous year.
- **Work Life Balance**: Categorical — Level of work-life balance.

2. **Random Undersampling:** Reduces the majority class size to balance data.
3. **Tomek Links:** Removes data points that are closest to another class to improve decision boundaries.
4. **NearMiss:** Selects data points from the majority class that are closest to the minority class.
5. **Borderline-SMOTE:** Generates synthetic samples near the decision boundary.
6. **ADASYN (Adaptive Synthetic Sampling):** Creates synthetic samples based on class density.
7. **Cluster Centroids:** Replaces samples with synthetic points using clustering methods.
8. **SMOTE-NC (Synthetic Minority Over-sampling for Nominal and Continuous):** A variation of SMOTE for mixed data types.
9. **MDO (Modified Distribution Over-sampling):** Creates synthetic samples based on probability distribution adjustments.
10. **Ensemble Learning-Based Methods:** Includes Balanced Random Forest and Easy Ensemble for better handling of imbalanced data.

4. Cost-Sensitive Learning for Imbalanced Data

Instead of oversampling, cost-sensitive methods were employed to adjust model performance by assigning different weights to misclassified samples.

- **Cost-Sensitive Logistic Regression:** Adjusts loss function to penalize misclassifications in the minority class.
- **Imbalanced Learning Approaches:**
  - Sampling Methods (SMOTE, Random Oversampling, etc.)
  - Cost-Sensitive Methods (Weighted Loss Functions)
  - Kernel-Based Methods (SVM with Weighted Kernel)
  - Active Learning Methods (Selective Sample Weighting)

5. Machine Learning Models

For predicting employee attrition, the following models were used:

- **Logistic Regression:** Provides interpretability in identifying key influencing factors.
- **Random Forest:** Handles complex interactions and provides higher accuracy.
- **Deep Learning (DNNs):** Captures non-linear patterns in attrition data.

Dataset

In this study, we utilized a dataset containing employee information to predict and analyze attrition patterns. The dataset was sourced from IBM HR Analytics and includes a wide variety of features that capture demographic, professional,

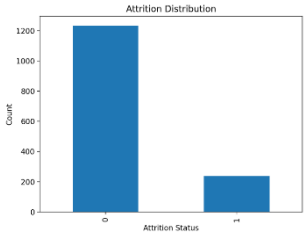
Reason: The model's ability to handle complex feature interactions (e.g., between atrrition and job satisfaction) and capture non-linear patterns in the data.

Exploratory Data Analysis (EDA)

The data revealed a **significant class imbalance**:

82% of employees stayed (No Attrition).

18% left the organization (Attrition).

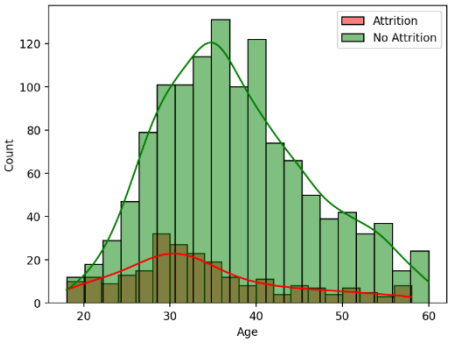


This imbalance necessitated the use of data balancing techniques (e.g., SMOTE) to improve model performance.

2. Age Distribution Before and After Attrition

Employees who left were concentrated in the **25–50 age group** (see next Figure 5).

The employees who remained were mostly elderly and young.

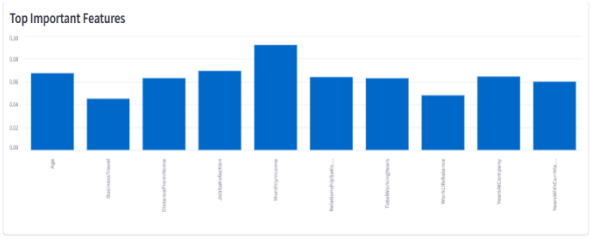


3. Attrition Rate by Department

The **Sales& human Resources** department had the highest attrition rate (**0.150)**

- **Years At Company**: Numeric (Discrete) — Number of years the employee has worked at the company.
- **Years Since Last Promotion**: Numeric (Discrete) — Number of years since the last promotion.
- **Years With Current Manager**: Numeric (Discrete) — Number of years working with the current manager.

This dataset, with its rich set of features, provides valuable insights into the factors influencing employee attrition. By analyzing and modeling these data points, we aim to develop a robust predictive framework to assist organizations in understanding and mitigating employee turnover.



Results and Discussion

In this section, we present the outcomes of the predictive models developed and discuss their implications for employee attrition management. The two primary models applied were Logistic Regression and Random Forest. Their performance was evaluated based on key metrics, including accuracy, precision, recall, F1-score, and AUC (Area Under Curve).

Results Overview

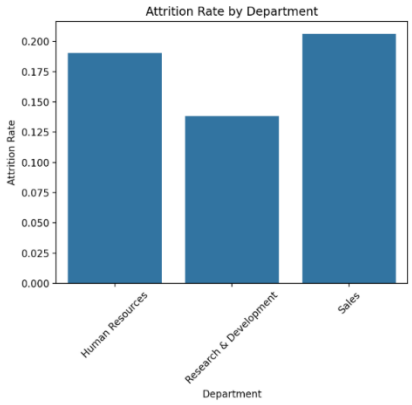
1. Model Performance

The **Random Forest** model outperformed other models (e.g., Logistic Regression), demonstrating superior predictive capabilities across all key metrics:

Metric	Random Forest
Accuracy	94.13%
Precision	**91.92%**
Recall	**96.76%**
F1-Score	**94.28%**
AUC	98.02%

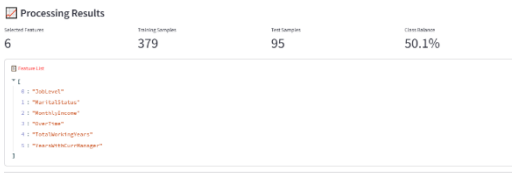
	precision	recall	f1-score	support
0	96.58%	91.50%	93.97%	24700.00%
1	91.92%	96.76%	94.28%	24700.00%
accuracy	94.13%	94.13%	94.13%	94.13%
macro avg	94.25%	94.13%	94.13%	49400.00%
weighted avg	94.25%	94.13%	94.13%	49400.00%

The **R&D** department had the lowest attrition rate (**0.150**) (see Figure 6).



2. Balancing with Random Under-Sampler

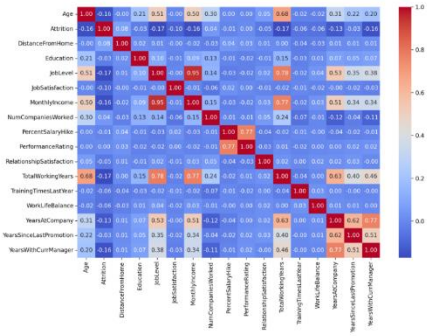
Selected Features: 6 features (e.g., Age, OverTime).
Data Size:
Training Samples: 379
Test Samples: 95
Class Balance: 50.1%



3. Balancing with SMOTE

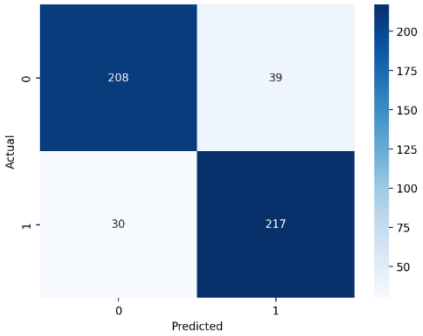
Selected Features: 6 features (e.g., JobLevel, YearsAtCompany).
Data Size:
Training Samples: 1,308
Test Samples: 162
Class Balance: 84.1%
(see Figure 4).

4. Correlation Matrix
Key correlations included:
Monthly Income and Job Satisfaction.
Overtime Hours and Stress Levels.



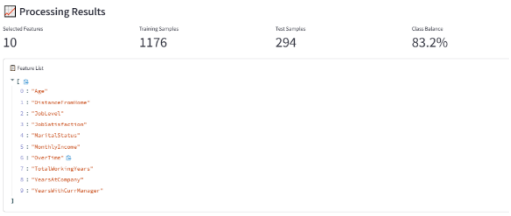
Model Evaluation

1. Confusion Matrix
True Positive (TP): 208 (correctly predicted attrition).
True Negative (TN): 217 (correctly predicted retention).
False Positive (FP): 39 (incorrectly predicted attrition).
False Negative (FN): 30 (incorrectly predicted retention).



Data Preprocessing Results

1. Balancing with Random Over-Sampler
Selected Features: 10 features (e.g., Age, MonthlyIncome, JobSatisfaction).
Data Size:
Training Samples: 1,176
Test Samples: 294
Class Balance: 83.2%

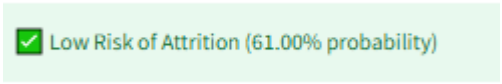


hybrid models and reinforcement learning for enhanced attrition prediction.

2. Key Influencing Factors

Monthly Income: Highest impact (0.09).
Job Satisfaction: Moderate impact (0.07).
Age: Moderate impact (0.07).

Prediction Result:



Implications for HR Management

1. **Targeted Retention Strategies:** Companies can focus on high-risk employees and design personalized intervention plans.
2. **Data-Driven Decision Making:** Insights derived from predictive analytics enable HR teams to base their decisions on evidence rather than intuition.
3. **Continuous Improvement:** Periodic model retraining with updated data ensures that predictions remain relevant.

Recommendations

1. Improve work-life balance through flexible scheduling.
2. Reassess compensation structures to ensure competitiveness.
3. Offer professional development opportunities.
4. Conduct regular employee engagement surveys to identify dissatisfaction early.

- **Conclusion**

This study highlights the power of predictive analytics in identifying factors influencing employee attrition. By leveraging machine learning, deep learning, and data balancing techniques, organizations can take proactive steps to improve employee retention and workplace satisfaction. Future research can explore

15. Kim, T. H., & Park, J. S. (2009). Do types of organizational culture matter in nurse job satisfaction and turnover intention? Leadership in Health Services, 22(1), 20-38

References:

1. Cost of Employee Attrition:
1. Cascio, W. F., & Boudreau, J. W. (2011). Investing in People: Financial Impact of Human Resource Initiatives. FT Press.
2. [Link](https://www.ftpress.com)
3. Factors Influencing Employee Attrition:
4. Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover. *Journal of Management*, 26(3), 463-488.
5. [DOI: 10.1177/014920630002600305](https://doi.org/10.1177/014920630002600305)
2. IBM HR Analytics Employee Attrition & Performance Dataset:
6. Kaggle. (n.d.). IBM HR Analytics Employee Attrition & Performance.
7. [Link](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset)
3. Machine Learning Models for Attrition Prediction:
8. Zhang, Z., & Zhang, Y. (2019). Predicting employee turnover using machine learning techniques. Journal of Big Data, 6(1), 1-20.
9. [DOI: 10.1186/s40537-019-0191-6](https://doi.org/10.1186/s40537-019-0191-6)
10. https://towardsdatascience.com/predicting-employee-turnover-7ab2b9ecf47e
11. https://hirebee.ai/blog/recruitment-metrics-and-analytics/predictive-analytics-for-employee-retention-forecasting-and-preventing-turnover/
12. Hassan, R. (2014). Factors influencing turnover intention among technical employees in information technology organization: A case of XYZ (M) Sdn. Bhd. International Journal of Arts and Commerce, 1(4), 53-63.
13. Lo, W. Y., Chien, L. Y., Hwang, F. M., Huang, N., & Chiou, S. T. (2017). From job stress to intention to leave among hospital nurses: A structural equation modelling approach. Journal of Advanced Nursing, 74(5), 677-688.
14. Mobley, W. H. (1977). Intermediate Linkages in the Relationship between Job Satisfaction and Employee Turnover. Journal of Applied Psychology, 62(2), 237-240.