International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 7 Issue 3, March 2026

Data Driven Attrition Prediction In Enterprise Organization

Rozina Naz[1], Nidhi Tijare[2], Shruti Lanjewar [3]
Science and Technology, G.H. Raisoni Skill Tech University Nagpur, Maharashtra, India[1]
Science and Technology, Sadabai Raisoni Women's College Nagpur, Maharashtra, India[2]
Science and Technology, Sadabai Raisoni Women's College Nagpur, Maharashtra, India[3]
Rozina.naz20@gmail.com[1], nidhi.tijare.bca@srwc.raisoni.net[2], shruti.lanjewar.bca@srwc.raisoni.net[3]

Abstract: *As employee attrition become a serious problem for companies as it affects productivity, disrupts work, and increases hiring costs. To come out of this problem, this paper brings a machine learning – based system which can predicts employee attrition using HR data. This system analyses historical employee records to identify trends such as is employee is willing to leave the company or voluntary resignation. The dataset should be clean and organized before applying the machine learning model, this also contain the algorithm such as logistic regression and random forest. This system works better than the traditional system as it predicts the risk of employee who is going to leave. performs better than traditional methods. This helps HR teams in planning effective retention strategies.*

Keywords: *Employee Attrition, HR Analytics, Machine Learning, Prediction Modelling, Workforce Retention.*

1. Introduction

Employee attrition means the employee is willing to leave or voluntary resignation. Attrition happens when employee leave a company because of resignation, retirement or dismissal. In this growing world, the company prefer to keep the talented employees in the organization. As there is a need of hiring employees it leads to increase in hiring cost which can cause loss of skilled knowledge and reduced motivation among remaining employees. Traditional HR systems mainly analyze past data to understand attrition trends, but they do not predict future turnover. As a result, these methods are reactive and cannot prevent employees from leaving. With the rise of Human Resource Information Systems (HRIS), organizations now have access to structured employee data. Forecast potential resignations in advance and analyze the employee attrition for that we need intelligent systems.

1.1 Research Problem

Various companies have employee data, but who might resign can't be predicted, for prediction they do not have strong predictive tools to detect. Initially there are system but there are some limitations such as limits feature usage, no detailed model comparison and poor consideration of practical use.

1.2 Objectives and Contributions

Main goals of our research include:

- Find attrition patterns to analyze the employee data.
- To develop a reliable prediction model.
- To compare multiple supervised learning techniques.
- To assist HR professionals in making data-driven decision.

DOI: https://doi.org/10.5281/zenodo.19666877

2. Literature Review

To maintain productivity and reduce workforce problems in companies for that prediction system plan very important role. Earlier, studies focused on statistical techniques such as correlation analysis and linear regression to understand the reason behind employee turnover. The complex nature of real HR datasets is not properly managed even for the basic analysis [1]. To advance the system machine learning, researchers began using supervised models to predict employee attrition. For simple and gives clear results we use Logistic Regression. Our studies show how that salary, job satisfaction, company experience, and work-life balance play a major role in employees leaving their jobs [2]. Logistic Regression works well in many cases, but it assumes a straight-line relationship between variable and outcomes, which is not always realistic in complex datasets. To handle the non-linear relationship and more interpretable inputs our system uses Decision Tree to make a clear decision, so it can give the result that how decision is made step by step [3]. When train from high-dimensional datasets, single tree models often suffer from overfitting and instability. To achieve higher accuracy and improve generalization and robustness, we use random forest as compared to individual classifier [4].

3. Proposed Methodology

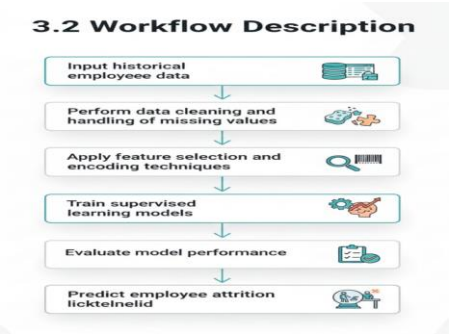
3.1 System Architecture

There five important modules:

1. Data input,
2. Preprocessing,
3. Feature engineering,
4. Model training and
5. Prediction output.

First it collects employee data and cleans it up. Then, it transforms that data into a format that works well with machine learning models.

3.2 Workflow description



3.3 Mathematical Model

The employee dataset is structured as:

$X = \{x_1, x_2, \ldots\ldots, x_n\}$
Let x_i denote the input feature vector for the i-th employee, with the binary target variable $y_i \in \{0, 1\}$ representing voluntary turnover ($y_i = 1$) or retention ($y_i = 0$):

$F(X) = P(y = 1 \mid X)$
We use different the models or that parameters are adjusted to minimize the following loss function:

$$L = - \sum [\, y \log (p) + (1 - y) \log (1 - p)$$

3.4 Flowchart Explanation

Firstly, you provide the dataset into the system. After that, the next step is preprocessing in that the data is sorted and cleaned and feature engineering is the machine learning model to get it ready. Once the data's prepared, you then train the data and test the in the model. The model basically learns from the training the data, then you check how well it does the test set basically used. In the end, you get the result that is attrition predictions and you can see how the model performed.

4. Implementation details

4.1 Hardware Requirements

The system does not need advanced or precision tackle for this study. Scholars and professionals can develop and test it on a standard computer that is generally available to them. The program can be run.

DOI: https://doi.org/10.5281/zenodo.19666877

The dataset used in the design can be handled by a computer with the stable processor, sufficient memory and acceptable storage.
Processor: Intel i3 or any original or advanced processor
RAM: Minimum 4 GB
Storage: At least 20 GB of free space

4.2 Software Requirements

- Operating System: Windows or Linux
- Programming Language: Python
- Tools: Jupyter Notebook, Anaconda
- Libraries: NumPy, Pandas, Matplotlib

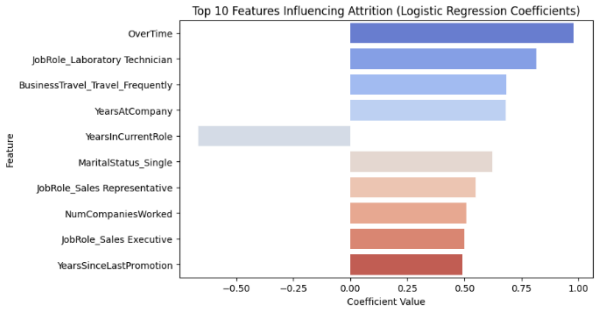
4.3 Dataset Description

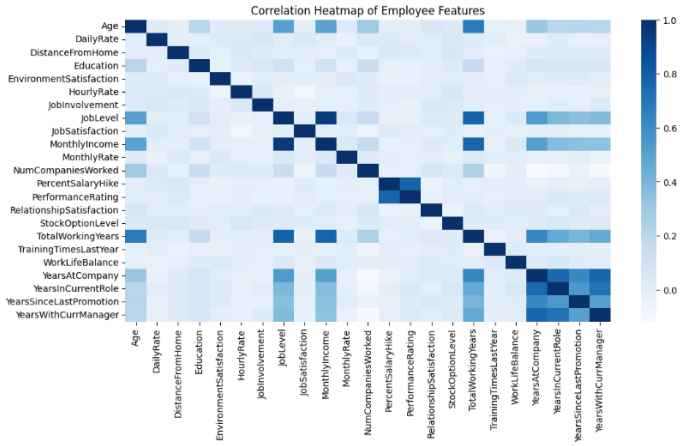
The study comprises of the IBM HR Analytics dataset having 1470 employee records and with 35 attributes. These attributes contribute and cover various different aspects which include demographics, job roles, compensation levels and performance metrics and many more.

5. Experimental Results and Discussion

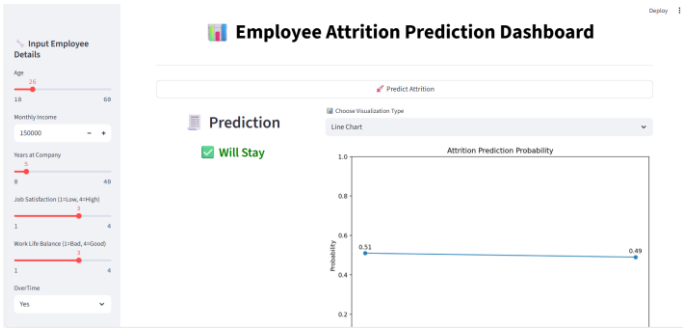
The system was evaluated and checked whether the model is working properly or not with accuracy, precision, recall and F1-score. Logistic regression was used as baseline because it provides good interpretability. Random forest has achieved higher accuracy than the logistic regression due to its potential to size complicated feature interplay effectively.

```
Logistic Regression Accuracy: 0.8945578231292517
Random Forest Accuracy: 0.8775510204081632

Classification Report (RF):
              precision    recall  f1-score   support

          No       0.88      1.00      0.93       255
         Yes       0.80      0.10      0.18        39

    accuracy                           0.88       294
   macro avg       0.84      0.55      0.56       294
weighted avg       0.87      0.88      0.83       294
```





Comparative analysis shows that ensemble-based model performed better than individual models. Visualization of using graphs and charts to highlights the contribution of key capabilities which include process job satisfaction. Monthly income and years spent in the company. The results shows that machine learning models can accurately predict employee attrition successfully and help HR department to take actions early.



DOI: https://doi.org/10.5281/zenodo.19666877

6. Conclusion and Future Work

This study introduced a system that uses the machine learning along with HR analytics to predict when employees might leave. The main goal was to carefully examine the employee records who are planning to leave. The simple idea was to carefully look into employee data and spot the signs that an employee maybe planning about quitting. Learning on data like this really improves their ability to detect employee attrition and tell how well companies can see attrition early, which helps them manage their teams better. If manager knows which employee may leave, they can actually do something about it. Manager can check if there is a problem at workplace and provide more support to the employee and offer promotion or growth opportunity. Instead of wasting resources on constant onboarding and recruitment, investing in retention maintains institutional knowledge and boosts efficiency. We can advance this system in future, as the system will uses the current data prediction could be more updated and accurate, so companies are not working on old information. As we will ask for employee feedback and survey answers, so are basically all those comments people leave and might reveal the hidden problems or real feelings of employee. And if they push into deep leaning, the predictions could get even sharper.

References

[1] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[2] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, Wiley, 2013.

[3] IBM Corporation, “IBM HR Analytics Employee Attrition Dataset,” 2016.

[4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Elsevier, 2011.

[5] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

DOI: https://doi.org/10.5281/zenodo.19666877