Employee Attrition Prediction Using Machine Learning: A Technical Report

Vidhaan Appaji, Anusha S. L. Bandaru and Rakesh R. Bhatija

Applied Artificial Intelligence, University of San Diego

AAI-500: Probability and Statistics for Artificial Intelligence

Dr. Zahid Hussain Wani

February 25, 2025

Abstract

Employee attrition is an important concern for organizations that seek to have a stable and productive workforce. This study investigates machine learning approaches to modeling employee attrition from data provided by a human resource database. The study focuses on data preprocessing, feature engineering, model selection, and evaluation to ensure maximum prediction performance. It uses the IBM HR Analytics Attrition Dataset from Kaggle (Pavan Subhasht, 2021) to develop and compare machine learning models for prediction of attrition rate of employees. The Random Forest model is used, as well as an improved version using Synthetic Minority Over-sampling Technique (SMOTE) to handle the class imbalances. The outcomes show that the basic model reaches 84.35% accuracy, while the balanced model enhances recall, demonstrating trade-offs between precision and recall.

Employee Attrition Prediction Using Machine Learning: A Technical Report

Introduction

Employee attrition has a substantial impact on the operations, productivity, and financial stability of business. Attrition prediction enables organizations to introduce proactively retention strategies. This research uses machine learning for identifying major drivers of employee attrition and developing prediction models based on Random Forest and SMOTE-improved methods. The objective of the research is to assess the quality of various models, recognize major attrition drivers, and propose data-based retention strategies.

Data Cleaning and Preparation

The dataset used for this research includes employee records that have multiple numerical and categorical variables. Preprocessing involves:

Handling Missing Values*.* A preliminary assessment indicated that the dataset has no missing values. However, redundant variables such as EmployeeCount, Over18, and StandardHours were eliminated as they do not contribute to predictive modeling.

Binary Encoding. Binary categorical variables like Attrition, Gender, and OverTime were converted into binary format.

One-Hot Encoding. Performed one-hot encoding on the categorical features such as businessTravel, Department, Education field, job role, and Marital Status to make them compatible with machine learning models.

Feature Selection. Removing unnecessary columns like EmployeeNumber, Over18, StandardHours, and EmployeeCount.

Class Imbalance Handling. Using SMOTE to enhance the minority class representation (Attrition = 1).

Feature Engineering. Feature importance shown in Figure 1 was analyzed to determine the most influential variables. Some key features affecting attrition included JobSatisfaction, WorkLifeBalance, OverTime, and MonthlyIncome.

Exploratory Data Analysis (EDA)

EDA was conducted to see variable distributions and correlations. Some of the key findings include:

- Employees working overtime have a higher likelihood of attrition.
- Job titles and education fields have differential attrition patterns.
- Lower satisfaction levels among employees result in more attrition.
- The Age variable indicates that younger employees have a higher attrition rate.
- Department-wise attrition patterns indicate higher turnover in sales jobs.

The visualization of histogram and bar plots were applied for verification purposes.

Model Selection

Two Random Forest models were used:

Baseline Random Forest Model*.* A typical Random Forest classifier that is trained on the original dataset.

SMOTE-Improved Random Forest Model. Synthetic Minority Over-sampling Technique (SMOTE) is an improved model of Random Forest Classifier that includes oversampling to address the class imbalance.

The second model aims to enhance recall while maintaining predictive stability.

Model Analysis

In measuring the performance of both models in predicting employee attrition, several performance metrics were used. Each metric gives a different insight into how accurately the model predicts employee attrition.

Accuracy. Accuracy is the ratio of correctly classified instances to all instances.

ROC-AUC Score. The Receiver Operating Characteristic - Area Under Curve (ROC-AUC) score is used to measure the classification model’s ability to differentiate between the classes. The score is 1 for perfect, 0.5 for random guessing.

Log Loss. Logarithmic Loss (Log Loss) measures how uncertain the model is when predicting. Lower is better.

Cohen’s Kappa. Adjusts for chance agreement and measures the agreement between predicted and actual values.Kappa value equal to 1.0 means perfect agreement, 0.5 means moderate agreement, 0.0 indicates No agreement beyond chance, negative Cohen’s Kappa indicates worse than chance.

R^2 Score. How well the model takes the target variable's variance into account (Attrition).

Precision (Class 1). Precision (Positive Predictive Value) measures the number of predicted positive cases (attrition cases) that are correct.

Recall (Class 1). Recall (Sensitivity) measures the number of attrition cases correctly identified.

Baseline Model Performance

The classification model predicted an accuracy of 84.35%, this means that it correctly classified 84.35% of employees as either staying or leaving. The model yielded an ROC-AUC of 0.7704 which suggests that there is a 77.04% probability that the model will assign a higher attrition risk score to an employee who actually leaves than to one who stays, indicating a moderate to strong predictive power. The log loss of 0.3756 was obtained which suggests that the model’s probability estimates are fairly good but could be improved.Kappa is obtained as 0.1113 that suggests slight agreement beyond chance. This means that while accuracy is high, the model might still be biased towards the majority class (employees who stay). However, this implication was overcome by the SMOTE-Improved model discussed further (see Table 1). An

R^2 score of 0.1442 is obtained that means that only 14.42% of the variance in employee attrition is explained by the model (see Appendix A for the overall summary of the results)

Table 1

Performance metrics of Random Forest and SMOTE-Improved Random Forest Models (see Appendix A for more).

Metric	Score of Baseline Random Forest Model	Score of SMOTE-Improved Random Forest Model
Accuracy	84.35%	80.27%
ROC-AUC Score	0.7704	0.7801
Log Loss	0.3756	0.4227
Cohen's Kappa	0.1113	0.4000
R2 Score	0.1442	0.1442
Precision (Class 1)	0.57	0.42
Recall (Class 1)	0.09	0.66

SMOTE-Improved Random Forest Model.

After applying SMOTE with a sampling strategy of 0.2, the performance of the improvised model was evaluated using multiple metrics. This model correctly classified 80.27% of employees as either staying or leaving. The model has a 78.01% probability of ranking an employee who leaves higher than an employee who stays. The model yielded a log loss of 0.4227 which is lower than the baseline model. A higher log loss suggests that while class balancing improved recall, it slightly reduced prediction confidence. The reason is that SMOTE introduces synthetic minority class samples, which can make the model more uncertain about classifications. Kappa obtained as 0.4000 indicates moderate agreement between predictions and actual values, a significant improvement from baseline Kappa which suggests the model makes better classifications beyond random chance for minority class (employees who leave).

SMOTE helps address the class imbalance, leading to better-balanced classification results. R^2 score is the same as the baseline random forest model. Precision is 42%. Precision decreased compared to the baseline, but this is expected because SMOTE increases recall at the cost of some false positives. Recall is 66% which indicates that the model correctly identifies 66% of employees who actually leave (see Appendix A for the overall summary of the results).

Recall is considered of high significance in the study because in HR analytics, recall (correctly identifying employees who leave) is often more important than precision. Identifying at-risk employees early allows companies to take preventive measures (e.g., retention programs, salary adjustments, promotions).

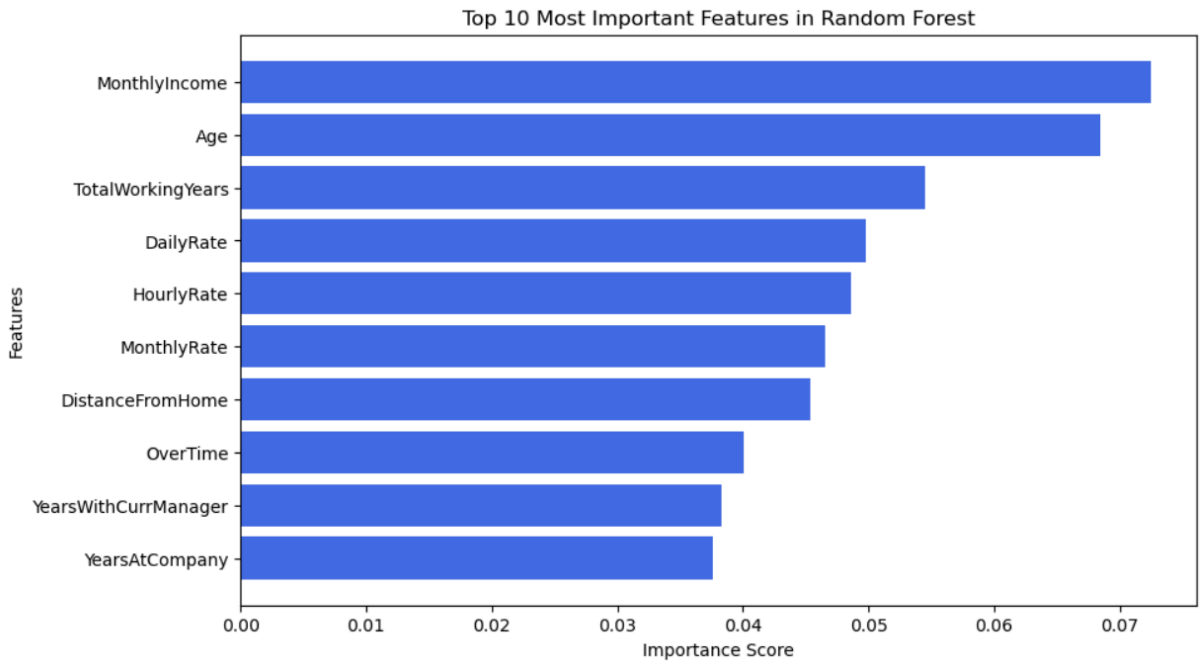
Table 1 shows the summary of the findings of both Random Forest and SMOTE-Improved Random Forest Model. Although the baseline model achieves high accuracy, the recall for predicting attrition cases is low. To mitigate class imbalance, SMOTE was applied with a sampling strategy of 0.2. The fine-tuned Random Forest model improved recall. This model demonstrates an improved recall, indicating better identification of attrition cases, albeit at a slight cost to accuracy.

Feature Importance Analysis

The top 10 ranked important features identified in the Random Forest model (see figure 1) include: Age, Monthly Income, Total Working Years, Daily Rate, Hourly Rate, Monthly Rate, Distance from Home, Over Time, Years with Current Manager, Years at Company.

Figure 1

Feature importances of Random Forest Model

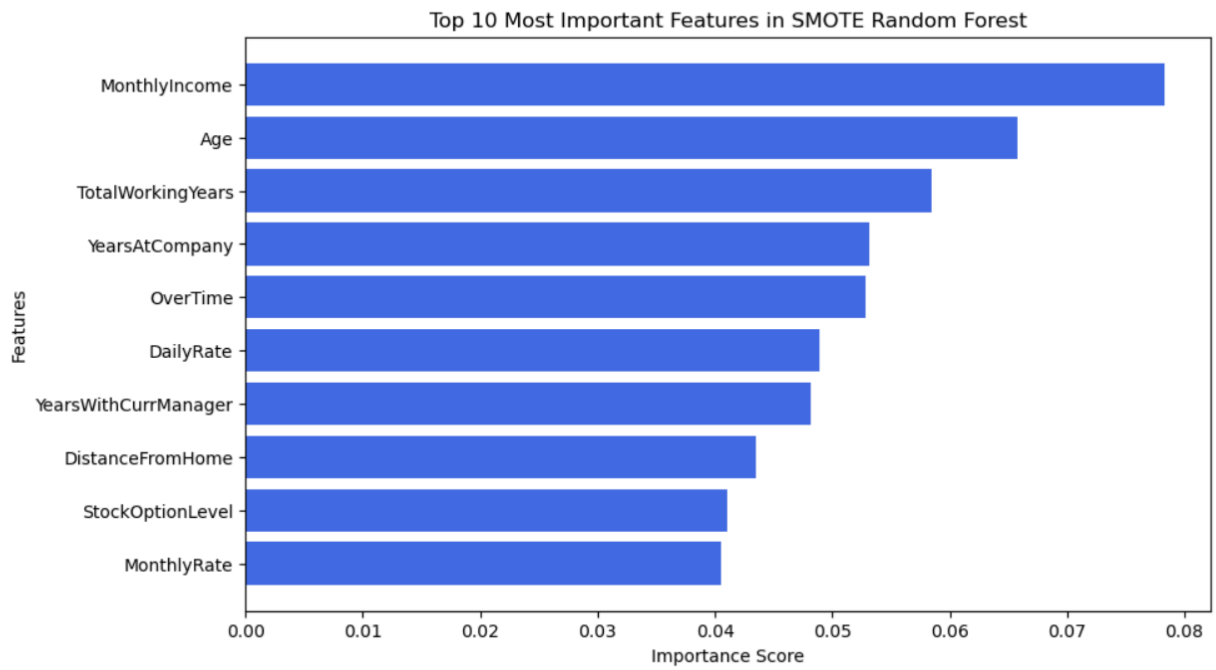


The top 10 most important features identified in the SMOTE Random Forest model (see figure 2) include: Monthly Income, Age, Total Working Years, Over Time, Daily Rate, Years with Current Manager, StockOption Level, Monthly Rate.

These findings align with organizational behavior research, reinforcing the importance of employee experience, compensation, and work-life balance in attrition.

Figure 2

Feature Importances of the SMOTE Random Forest Model



Conclusion and Recommendations

The study proves that machine learning methods, specifically Random Forest with SMOTE, is useful to predict employee attrition rate with high accuracy. The main conclusions are:

- The baseline Random Forest model is highly accurate but lacks recall.
- The SMOTE-augmented model has better recall, making it ideal for anticipatory HR action.
- OverTime, Job Role, and Monthly Income are significant predictors of attrition.

Recommendations for Organizations

- Adopt flexible work arrangements to mitigate overtime issues.
- Implement high-risk job-role targeted retention schemes.
- Provide fair compensation and opportunities for career progression.

- Regularly conduct employee satisfaction surveys to analyze the mood within the workplace.

Organizations are able to make informed HR choices and minimize turnover rates by making use of predictive analytics.

References

Agresti, A., & Kateri, M. (2020). *Foundations of statistics for data scientists: A comprehensive approach*. CRC Press.

Subhasht, P. (n.d.). *IBM HR Analytics Employee Attrition & Performance Dataset* . Kaggle. Retrieved from https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Appendix A

Summary of the results

Figure A1

Performance metrics of Random Forest Model

```
==== Model Performance Metrics ====
Accuracy: 0.8435
ROC-AUC Score: 0.7704
Log Loss: 0.3756
Cohen's Kappa Score: 0.1113
R² Score (on Probabilities): 0.1442

==== Classification Report ====
              precision    recall  f1-score   support

           0       0.85      0.99      0.91       247
           1       0.57      0.09      0.15        47

    accuracy                           0.84       294
   macro avg       0.71      0.54      0.53       294
weighted avg       0.81      0.84      0.79       294


Model saved successfully!

==== Feature Importance (Top 10) ====
                 Feature  Importance
10         MonthlyIncome    0.072520
0                    Age    0.068437
18     TotalWorkingYears    0.054460
1              DailyRate    0.049831
6             HourlyRate    0.048613
11           MonthlyRate    0.046555
2       DistanceFromHome    0.045405
13              OverTime    0.040045
24  YearsWithCurrManager    0.038292
21        YearsAtCompany    0.037577
```

Figure A2

Performance metrics of Random Forest Model

```
==== Model Performance Metrics ====
Accuracy: 0.8027
ROC-AUC Score: 0.7801
Log Loss: 0.4227
Cohen's Kappa Score: 0.4000
R² Score (on Probabilities): 0.1442

==== Classification Report ====
              precision    recall  f1-score   support

           0       0.93      0.83      0.88       247
           1       0.42      0.66      0.52        47

    accuracy                           0.80       294
   macro avg       0.68      0.74      0.70       294
weighted avg       0.85      0.80      0.82       294


==== Feature Importance (Top 10) ====
                  Feature  Importance
10          MonthlyIncome    0.078343
0                     Age    0.065817
18      TotalWorkingYears    0.058422
21         YearsAtCompany    0.053113
13               OverTime    0.052789
1               DailyRate    0.048934
24   YearsWithCurrManager    0.048124
2        DistanceFromHome    0.043460
17       StockOptionLevel    0.041040
11            MonthlyRate    0.040558
```

Table A1

Comparison of Baseline vs. SMOTE-Enhanced Model

Metric	Score of Baseline Random Forest Model	Score of SMOTE-Improved Random Forest Model	Change
Accuracy	84.35%	80.27%	Decreased
ROC-AUC Score	0.7704	0.7801	Improved
Log Loss	0.3756	0.4227	Increased
Cohen's Kappa	0.1113	0.4000	Significant Improvement
R2 Score	0.1442	0.1442	No change
Precision (Class 1)	0.57	0.42	Decreased
Recall (Class 1)	0.09	0.66	Improved

Figure A3

Model Comparison Graph

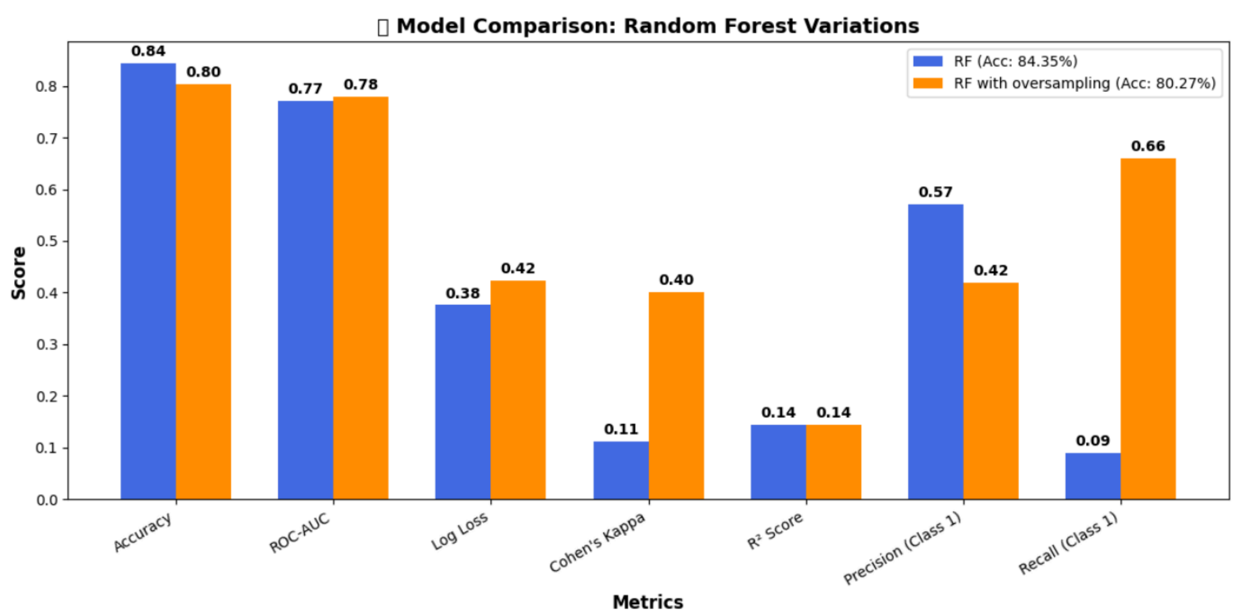


Figure A4

Console Application of the Random Forest Model

```
Choose Model (RF for Random Forest / RFO for Random forest oversampled): RF

Enter Employee Details for Prediction ◆

Age: 35
Distance From Home: 5
Job Level (1-5): 3
Monthly Income: 8000
Total Working Years: 10
Years at Company: 6
Daily Rate: 373

Select Department:
Human Resources
Research & Development
Sales
ter the number corresponding to the department: 2

Select Education Field:
Human Resources
Life Sciences
Marketing
Medical
Other
Technical Degree
ter the number corresponding to the education field: 2

Education Level (1-5): 4

Prediction Result ◆
Model Used: Random Forest
Employee is **likely to stay**. (Probability: 0.45)
```

Figure A5

Console Application of the SMOTE-Improved Random Forest Model

```
 Choose Model (RF for Random Forest / RFO for Random forest oversampled): RFO

 Enter Employee Details for Prediction ◆

 Age: 35
 Distance From Home: 5
 Job Level (1-5): 3
 Monthly Income: 8000
 Total Working Years: 10
 Years at Company: 6
 Daily Rate: 373

 Select Department:
. Human Resources
. Research & Development
. Sales
nter the number corresponding to the department: 2

 Select Education Field:
. Human Resources
. Life Sciences
. Marketing
. Medical
. Other
. Technical Degree
nter the number corresponding to the education field: 2

 Education Level (1-5): 4

 Prediction Result ◆
 Model Used: Random Forest Oversampled
 Employee is **likely to stay**. (Probability: 0.39)
```

Appendix B

Repository Information

The full source code used in this project is available at the following repository:

Appaji, V., Bandaru, A., & Bhatija, R. (2024). Project-AAI-500 [Source code]. GitHub. https://github.com/rbhatija/Project-AAI-500

This repository contains all scripts for data preprocessing, analysis, and visualization. Users can access the latest updates and documentation in the repository's README file.