Employee Attrition Prediction Using Machine Learning Models

Student Name: Vishnu Narayanan Harish

Class: M.Tech AML

Institution: Amrita Viswa Vidya Peetham, Amritapuri

Faculty Mentor: Dr. Swaminathan

Abstract

Employee attrition remains a major challenge for organizations, particularly in the information technology sector where skilled labor mobility is high. The goal of this study is to predict employee attrition using multiple machine learning models and identify the most effective classifier. The dataset, derived from an IT workforce, was preprocessed with one-hot encoding and balanced using the SMOTE technique to address class imbalance. Logistic Regression, Decision Tree, Random Forest, and XGBoost models were trained and evaluated. The Random Forest model achieved the highest accuracy, providing both predictive insights and feature importance analysis to assist HR departments in proactive retention planning.

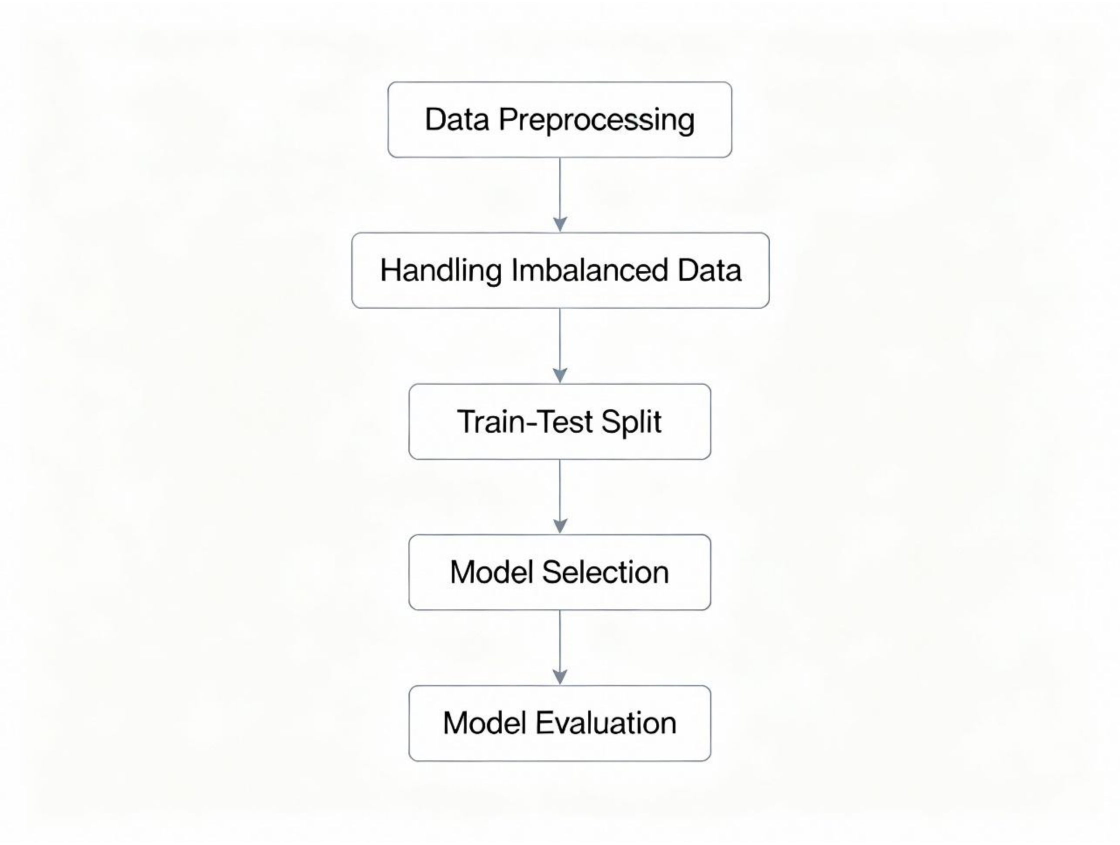
Introduction

Employee attrition, the process where employees voluntarily or involuntarily leave an organization, has significant financial and operational implications. Predicting attrition allows management to identify at-risk employees early and implement effective retention strategies. However, employee attrition data often exhibits high dimensionality with categorical variables (such as job role, department, overtime) and severe class imbalance (fewer cases of attrition). Traditional statistical methods struggle to handle these complexities. This project aims to overcome these limitations using machine learning models capable of capturing nonlinear relationships and handling imbalanced datasets.

Problem Statement

Organizations struggle to identify employees who may leave in the near future due to lack of reliable predictive tools. There is a need for a data-driven approach that analyzes key employee factors and predicts attrition accurately.

Methodology



1. Data Preprocessing

- Read CSV dataset using pandas.
- Target variable: Attrition (mapped as 1 for "Yes" and 0 for "No").
- Categorical columns were identified and encoded using OneHotEncoder

2. Handling Imbalanced Data

- Used Synthetic Minority Oversampling Technique (SMOTE) to balance the positive (leaving) and negative (non-leaving) classes.

3. **Model Selection**
    - Machine learning models used:
        - Logistic Regression
        - Decision Tree Classifier
        - Random Forest Classifier
        - XGBoost Classifier

4. **Model Training**

    - A Random Forest Classifier with 300 estimators was trained to identify complex patterns in the data and predict the crime type accurately.

5. **Model Evaluation**

    - Evaluated using Accuracy, Precision, Recall, and F1-score.
    - Plotted accuracy comparison and confusion matrix for the best-performing model
    - Computed feature importances to interpret model behaviour.

Dataset

Source: Custom synthesized dataset representing IT employee behaviour

Total Records: 1000

Features (X): Age, Gender, Department, Education, JobRole, MonthlyIncome, DistanceFromHome, YearsAtCompany, JobSatisfaction, WorkLifeBalance, OverTime

Target (Y): Attrition (1 = Yes, 0 = No)

Train-Test Split: 75% training, 25% testing

Implementation

Tools and Libraries

- Python
- pandas, numpy
- scikit-learn
- xgboost
- seaborn, matplotlib
- imbalanced-learn (for SMOTE)

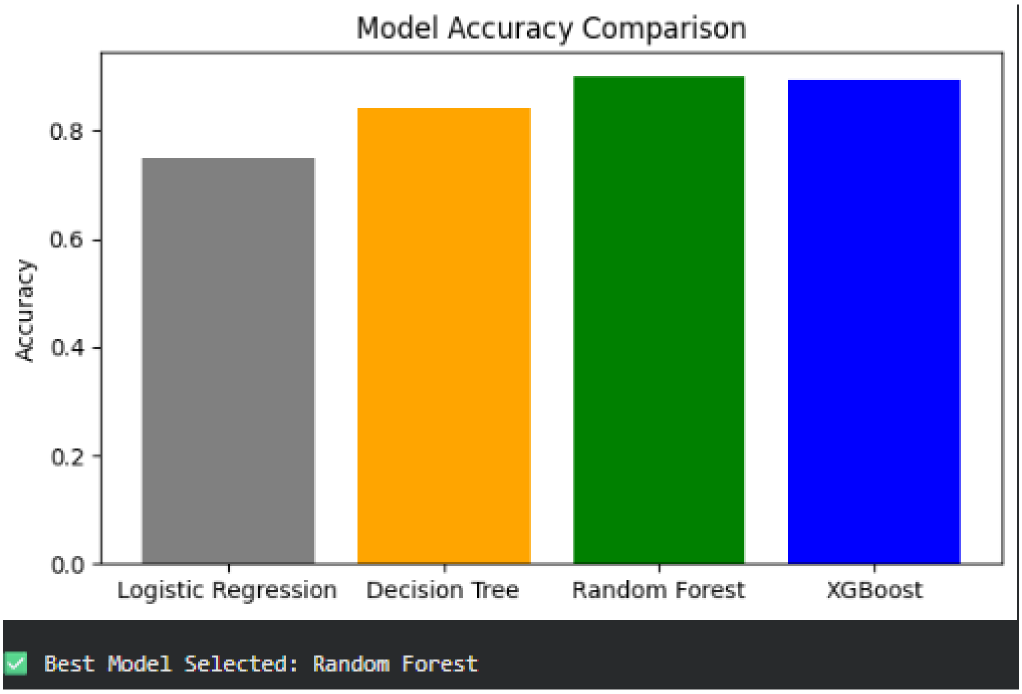
Algorithms Used

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost

GitHub:https://github.com/vishnunarayananharish/Employee-Attrition-Prediction-Using-Machine-Learning-Models

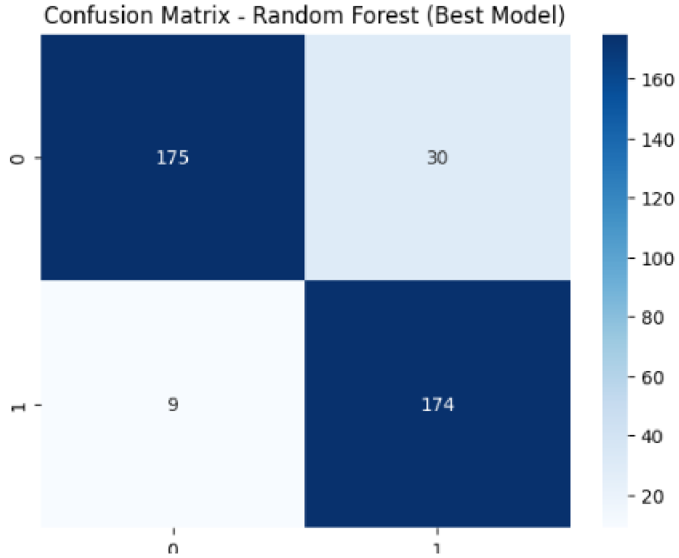
Results

The Random Forest Classifier acquired **89.9%** accuracy while, XGBoost achieved an overall accuracy of **89.4%.**



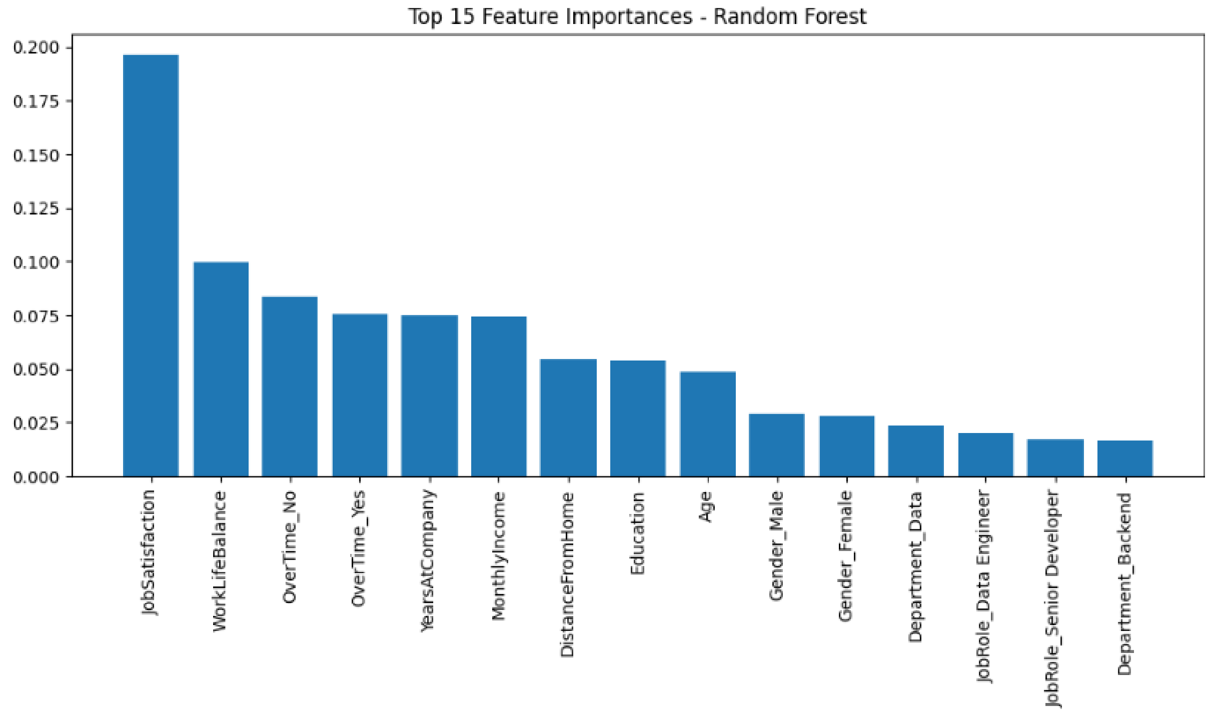
```
✅ Best Model Selected: Random Forest
```

The confusion matrix for Random Forest showed strong classification performance with minimal false negatives, suggesting reliable detection of potential leavers



Feature Importance Plot: Highlights which features influenced the predictions most

---



Prediction Example

Example Input:

Age: 29

Gender: Male

Department: Backend

Education: 3

JobRole: Software Engineer

MonthlyIncome: 45000

DistanceFromHome: 12

YearsAtCompany: 1

JobSatisfaction: 2

WorkLifeBalance: 2

OverTime: Yes

Predicted Output: LIKELY to Leave

Conclusion

This project demonstrated the effective application of machine learning techniques to predict employee attrition in an IT organization. The comparative study indicated that tree-based models, especially Random Forest, outperform linear baselines like Logistic Regression. Balancing the dataset with SMOTE significantly improved prediction fairness across classes. Overall, the model can serve as a decision-support tool for HR departments, identifying at-risk employees early and enabling proactive intervention strategies.

References

- Scikit-learn Documentation
- XGBoost Documentation
- imbalanced-learn Documentation
- HR Attrition Study Datasets and Research