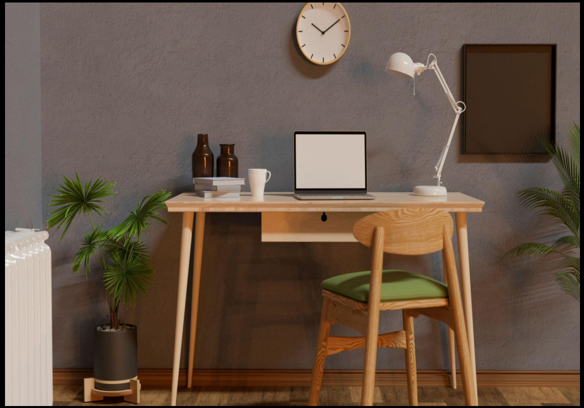
Employee Attrition Prediction



Alekhya Voleti

Machine Learning Fall 2025

Introduction & Problem Formulation

Background

Employee attrition is a major challenge across industries

High turnover increases recruitment, training, and onboarding costs

Loss of experienced employees impacts productivity and team morale

Problem Statement

Organizations need a way to predict attrition before it happens

This project formulates attrition as a binary classification problem

Machine Learning Objective:

Predict whether an employee is likely to leave based on historical employee data

Dataset Overview

IBM HR Analytics Employee Dataset (Kaggle)

Dataset Details

- 1,470 employee records
- 35 features
- Mix of demographic, job-related, and satisfaction variables

Target Variable

- Attrition (Yes/No)

Why this Dataset

- Real-world HR Data
- Suitable supervised classification tasks

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	RelationshipSatisfaction
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...	1
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...	4
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	2
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	3
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	4

Hours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
80	0	8	0	1	6	4	0	5
80	1	10	3	3	10	7	1	7
80	0	7	3	3	0	0	0	0
80	0	8	3	3	8	7	3	0
80	1	6	3	3	2	2	2	2

PROJECT OBJECTIVES

This project applies supervised machine learning techniques to predict employee attrition using real-world HR data. The objectives focus on building accurate predictive models, comparing alternative approaches, and extracting interpretable insights that can support data-driven employee retention strategies.

01 Predict Employee Attrition

Develop a machine learning model that predicts whether an employee is likely to leave a job based on historical demographic, job-related, and satisfaction data.

---

02 Compare Multiple ML Models

Train and evaluate multiple classification algorithms, including Logistic Regression, Random Forest, and XGBoost, to compare performance across standard evaluation metrics.

---

03 Interpret Results & Generate Insights

Analyze feature importance and model explanations to identify key factors contributing to employee attrition.

Approach

Workflow

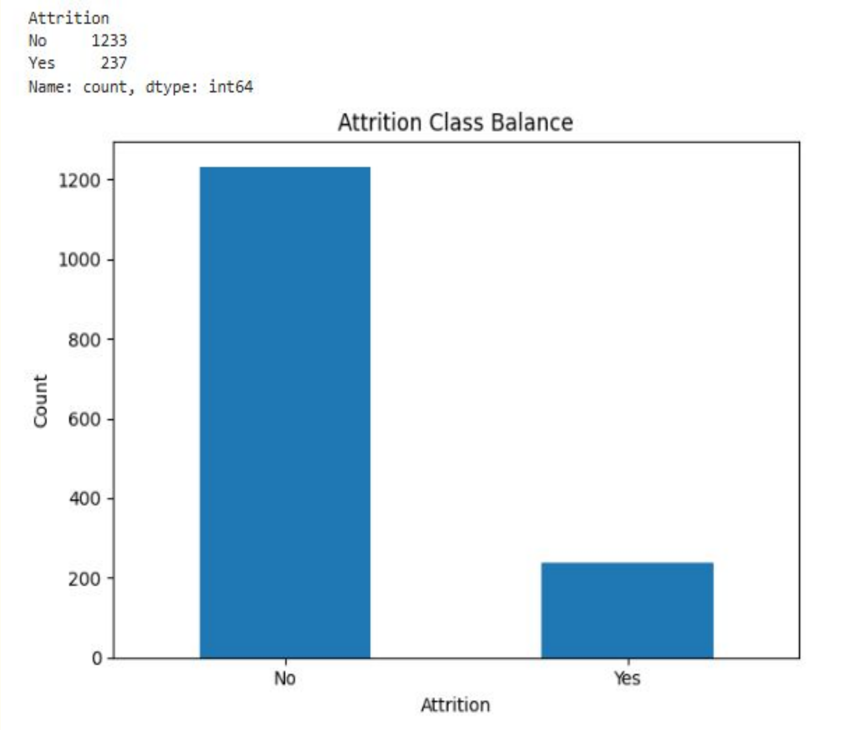
1. Data loading and cleaning using Python
2. Exploratory Data Analysis (EDA)
3. Data Preprocessing
   a. Encode categorical variables
   b. Scale numerical features
   c. Address class imbalance (SMOTE)
4. Model training and comparison

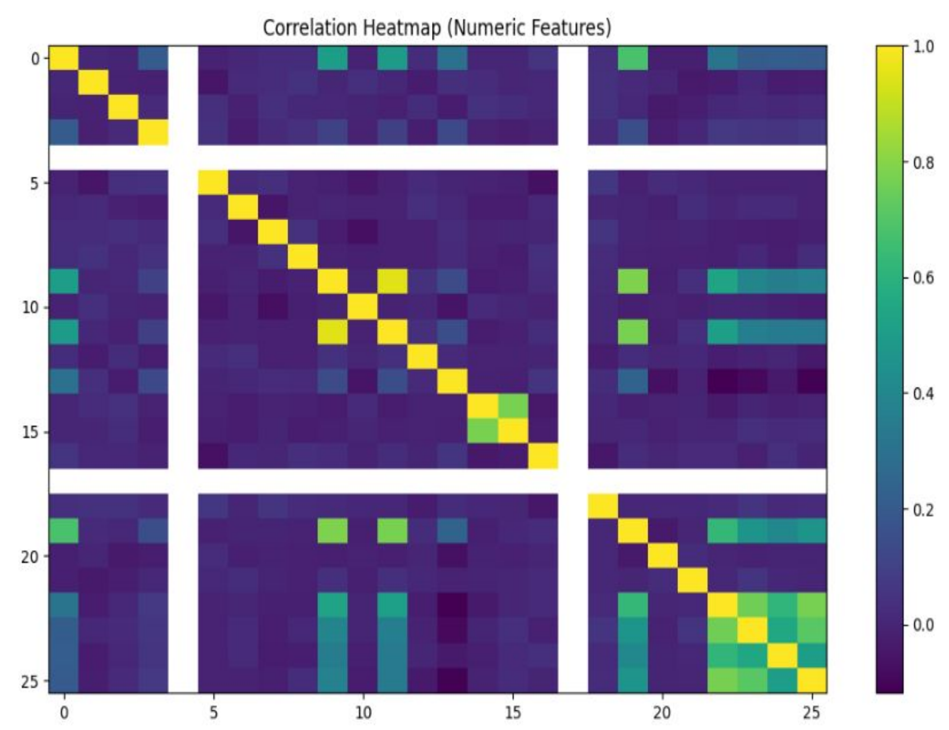
Models

- ***Logistic Regression***
- ***Random Forest***
- ***XGBoost***

Data → EDA → Preprocessing → Model Training → Evaluation

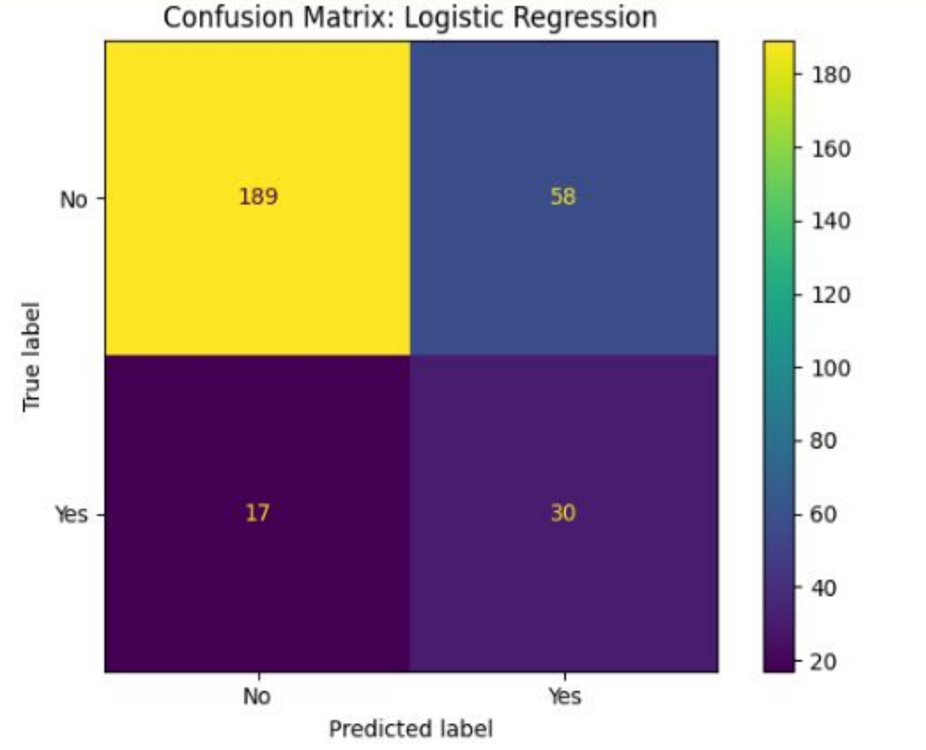
Preprocessing

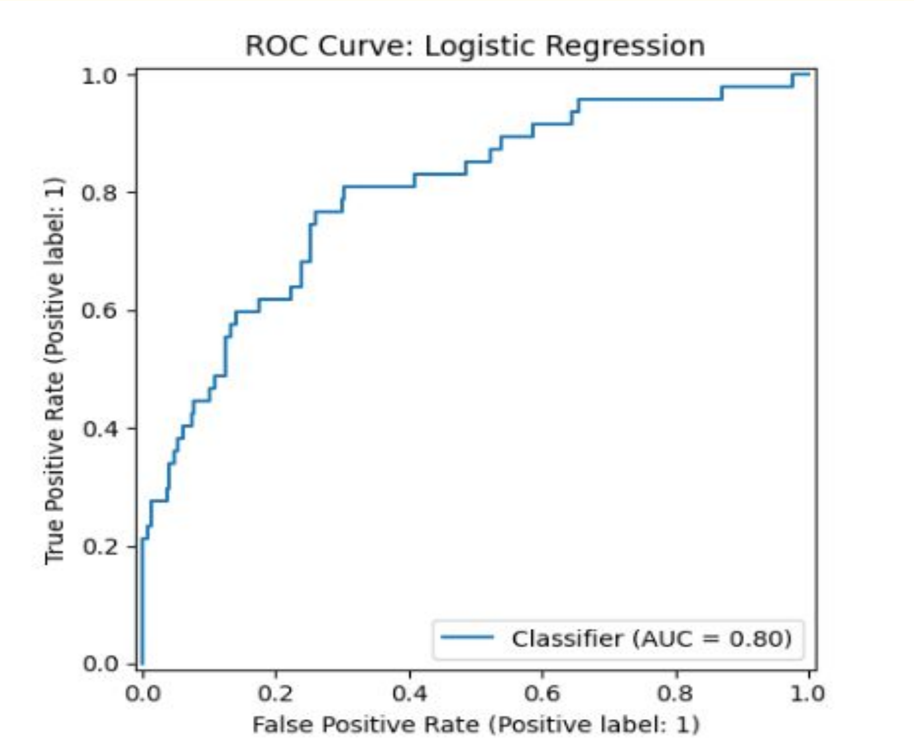




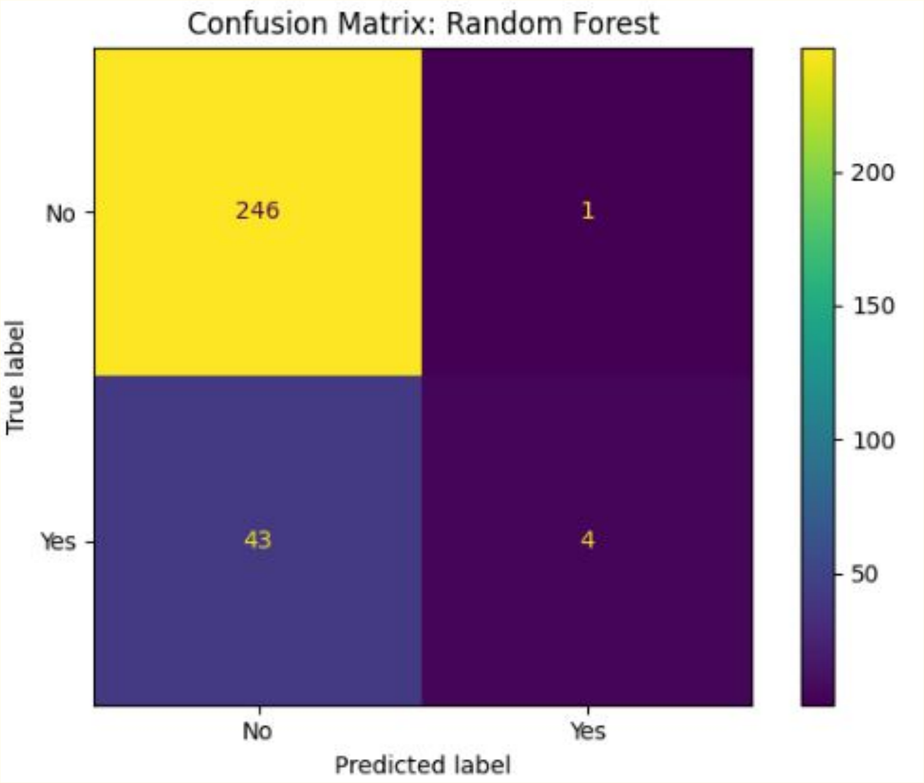
Model Evaluation & Comparisons

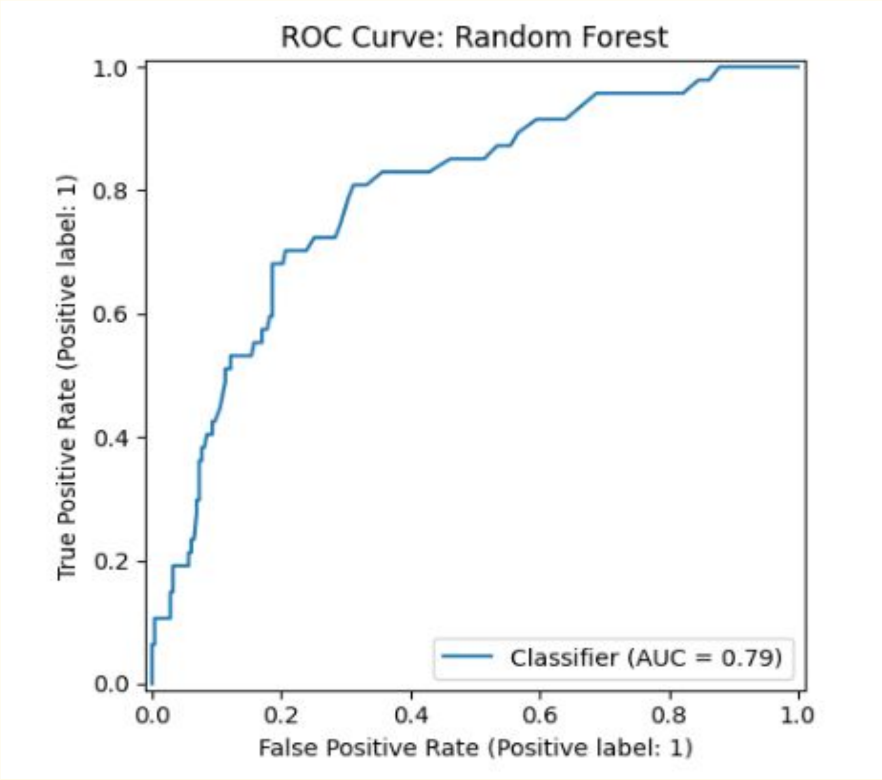
Logistic Regression



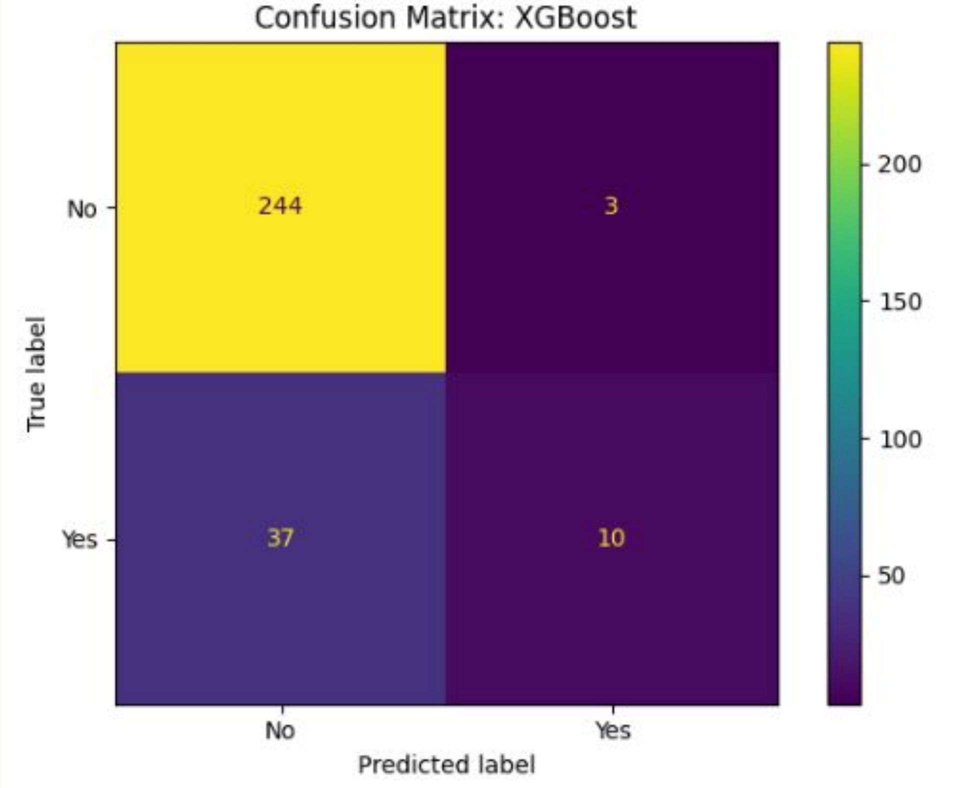


Random Forest

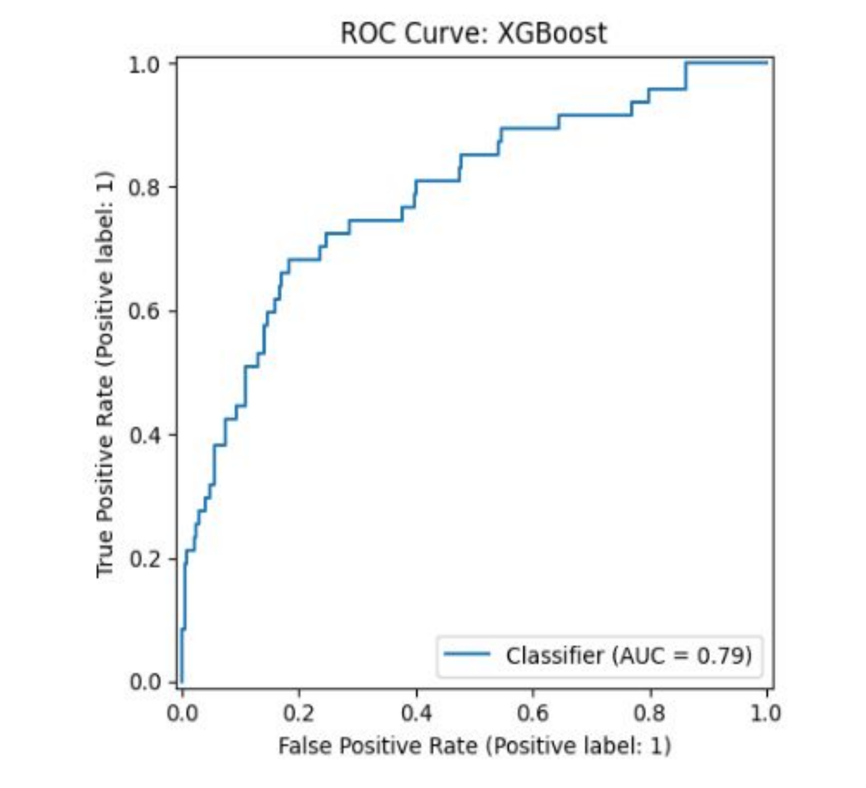




XGBoost



Confusion Matrix: XGBoost
244
3
37
10
No
Yes
True label
Predicted label
200
150
100
50



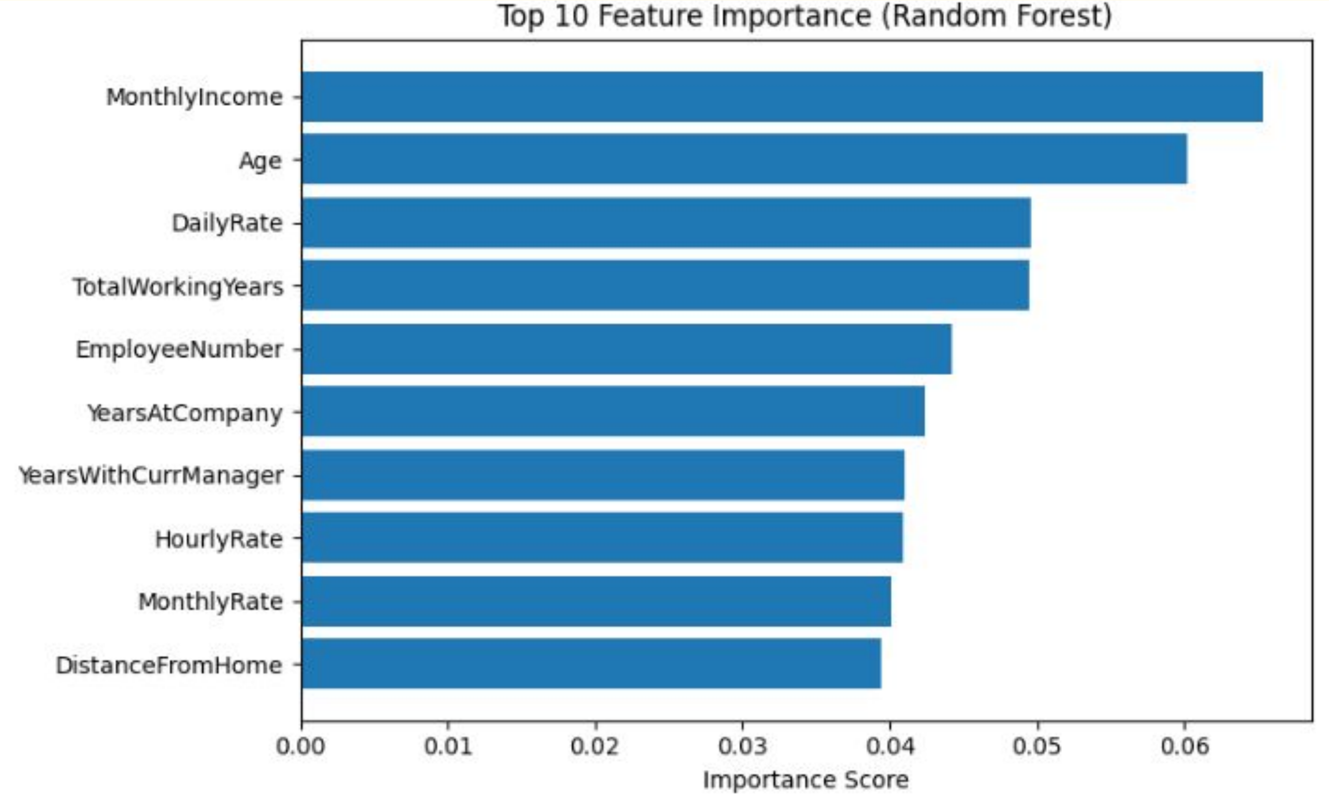
ROC Curve: XGBoost
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
Classifier (AUC = 0.79)
0.0
0.2
0.4
0.6
0.8
1.0

Compared Results

	Model	Accuracy	Precision	Recall	F1	ROC_AUC
0	Logistic Regression	0.744898	0.340909	0.638298	0.444444	0.797829
1	Random Forest	0.850340	0.800000	0.085106	0.153846	0.790335
2	XGBoost	0.863946	0.769231	0.212766	0.333333	0.785511

Feature Importance

Top Features & Interpretations



Feature	Importance
MonthlyIncome	0.065387
Age	0.060171
DailyRate	0.049635
TotalWorkingYears	0.049479
EmployeeNumber	0.044250
YearsAtCompany	0.042412
YearsWithCurrManager	0.041030
HourlyRate	0.040927
MonthlyRate	0.040143
DistanceFromHome	0.039446
NumCompaniesWorked	0.035097
OverTime_No	0.031430
StockOptionLevel	0.030501
YearsInCurrentRole	0.028155
PercentSalaryHike	0.028084

Business Insights & Recommendations

Model interpretation shows that attrition is driven by a combination of workload, compensation, satisfaction, and tenure-related factors rather than a single cause. These results support targeted, data-driven retention strategies. Key recommendations include prioritizing early-tenure employees, monitoring overtime as a leading risk indicator, and aligning compensation and engagement initiatives with high-risk employee segments to proactively reduce turnover.

Key Drivers of Employee Attrition

- **Overtime:** *Employees working overtime have a significantly higher likelihood of attrition, indicating workload and burnout as key drivers.*
- **Monthly Income:** Lower compensation is strongly associated with higher attrition risk, highlighting the impact of pay dissatisfaction.
- **Job Satisfaction:** Employees with lower job satisfaction scores are more likely to leave, emphasizing the importance of engagement and role alignment.
- **Tenure:** Attrition is more common among employees with shorter tenure, suggesting early-stage retention is especially critical.
- **Work-Life Balance:** Poor work-life balance correlates with higher attrition, reinforcing the need for flexible and supportive work policies.

Conclusion

- Employee attrition was successfully modeled as a **binary classification problem** using supervised machine learning
- **XGBoost achieved the best overall performance**, with ROC-AUC ≈ **0.80** and accuracy **>85%**, demonstrating strong predictive capability
- Logistic Regression provided **higher recall for attrition cases**, highlighting a trade-off between overall accuracy and early risk detection
- Model interpretation confirmed that attrition is **multifactorial**, driven by workload, compensation, satisfaction, tenure, and work-life balance
- Results support **targeted, data-driven retention strategies** rather than one-size-fits-all HR policies

Thank you!