# IBM HR Employee Attrition Report

IBM HR Employee Attrition Report

## Project Overview

This project explores employee attrition using the IBM HR dataset. It applies various data science techniques including exploratory data analysis (EDA), feature transformation, and supervised classification models to predict the likelihood of attrition. The main objective is to empower HR teams with data-driven insights to proactively identify high-risk employees and reduce turnover.

## Dataset Description

The dataset contains 35 columns and 1470 entries. Key attributes include:

- Age: Age of the employee

- Attrition: Whether the employee left (Yes/No)

- BusinessTravel: Frequency of travel (e.g., Travel\_Rarely, Travel\_Frequently)

- Department: Employee department (e.g., Sales, R&D, HR)

- JobRole: Job title (e.g., Sales Executive, Lab Technician)

- MonthlyIncome: Monthly salary

- OverTime: Whether the employee worked overtime

- YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion: Experience indicators

## Exploratory Data Analysis (EDA)

EDA revealed key patterns:

- Employees who left had notably lower satisfaction, income, and work-life balance scores.

- High attrition rates were observed in roles such as Laboratory Technician and Sales Executive.

- Overtime work was strongly correlated with higher attrition.

Visualizations were used to support findings, including bar plots and heatmaps highlighting feature correlations.

## Model Evaluation and Comparison

Several models were tested:

- Logistic Regression (with threshold tuning): Achieved the most balanced performance, improving minority class recall.

- LDA: Performed similarly post threshold adjustment.

- KNN: Showed signs of overfitting and poor recall on attrition class.

- QDA: Struggled due to class imbalance and non-linear separation.

Recall scores before and after threshold optimization were benchmarked. Logistic Regression at threshold 0.35 provided recall improvements while maintaining acceptable precision.

## Model Insights and Explainability

SHAP was used to interpret model decisions. Top features influencing attrition included OverTime, JobRole, MonthlyIncome, and WorkLifeBalance. SHAP force plots helped visualize how specific feature values contribute to individual predictions.

## Recommendations

- Adopt Logistic Regression at threshold 0.35 for balanced precision and recall.

- Target retention strategies toward employees working Overtime and in high-risk roles.

- Use SHAP visualizations to communicate drivers of attrition to HR.

- Consider using SMOTE or resampling to balance classes in future model iterations.

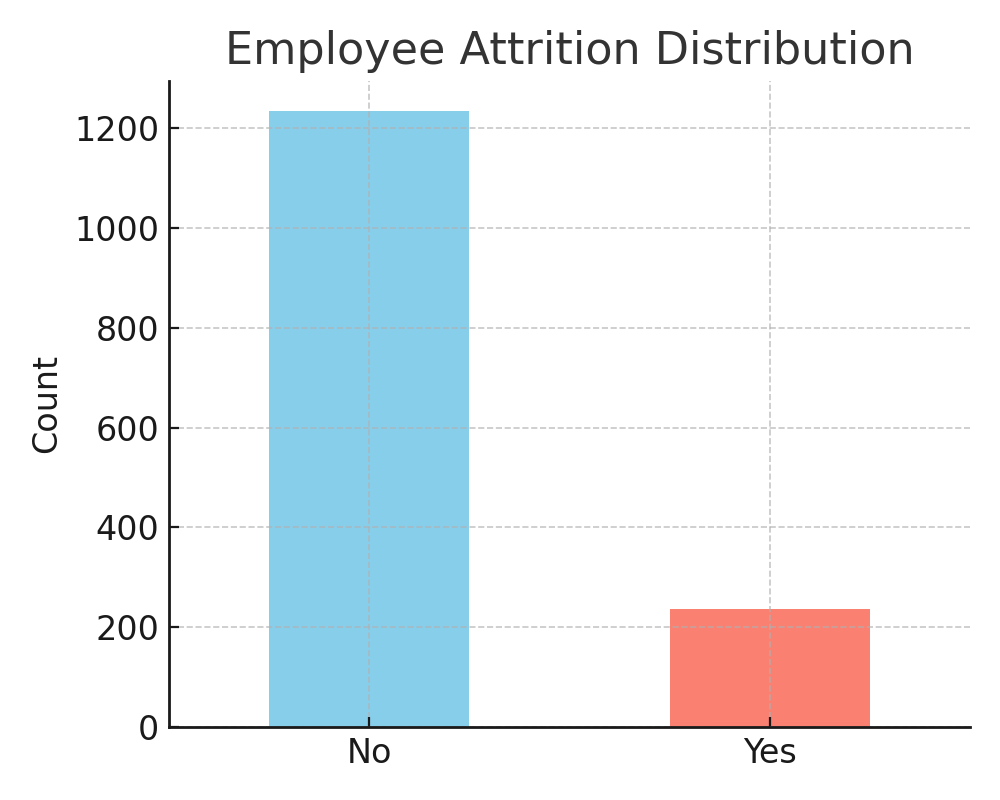
- Track performance over time to evaluate drift and adapt interventions.

This analysis provides actionable intelligence for employee engagement and strategic workforce planning.

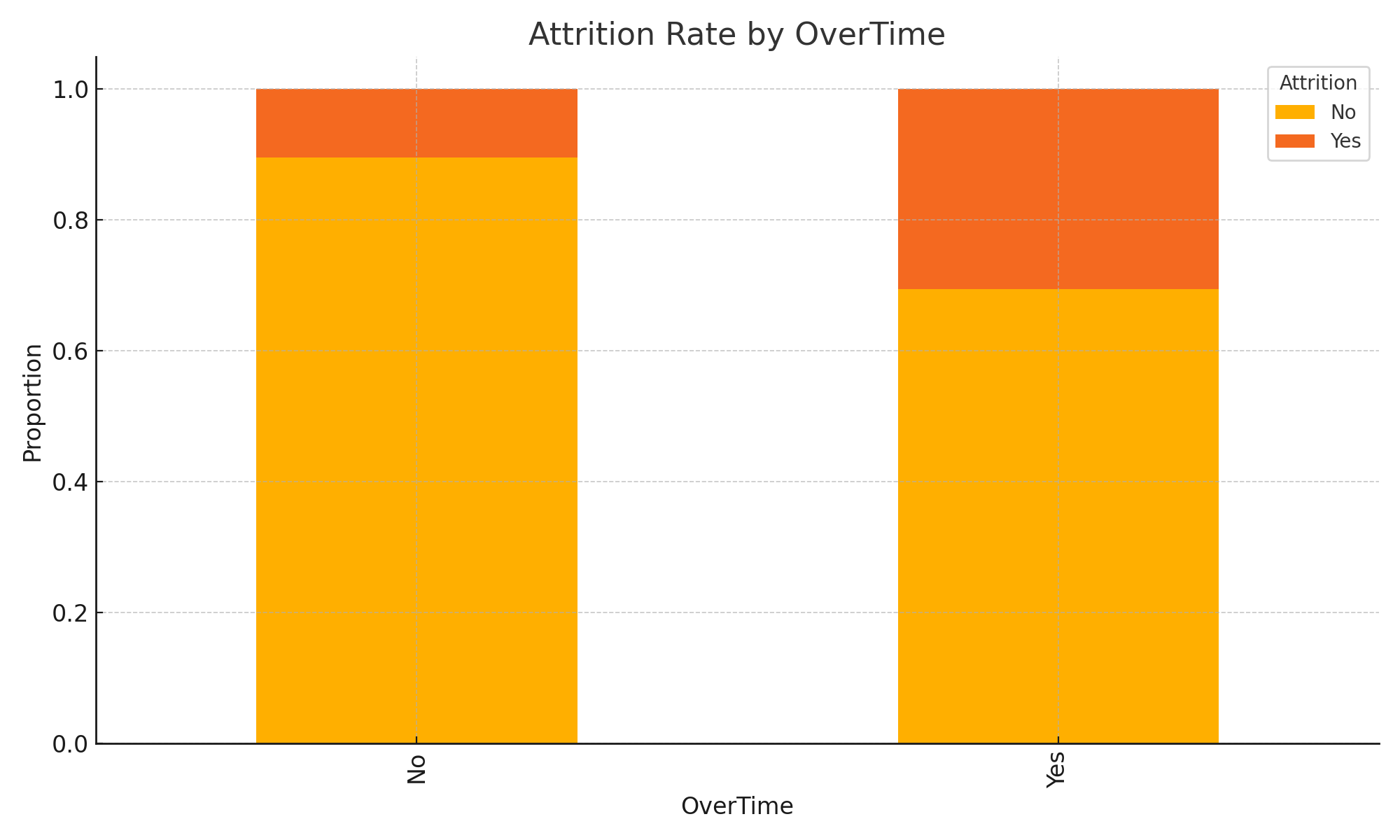
# 📊 Visualizations

The following visualizations support the exploratory data analysis and highlight key attrition patterns.

1. Employee Attrition Distribution:



2. Attrition Rate by OverTime:



3. Monthly Income Distribution by Attrition:

