EMPLOYEE ATTRITION PREDICTION – REPORT

Title: Predicting Employee Attrition using Machine Learning

Author: Ahsan Khizar

Date: 23/06/2025

1. Dataset Description

The dataset used in this project is the **IBM HR Analytics Employee Attrition Dataset**, which contains data on 1,470 employees. Each record includes demographic, job-related, and performance-related attributes.

Target Variable: Attrition – Whether an employee has left the company (Yes or No)

Key Features:

- Age, Gender, Job Role
- Monthly Income
- Overtime
- Job Satisfaction
- Work Life Balance

2. Data Preprocessing

- Converted categorical variables into numeric using **one-hot encoding** (pd.get_dummies).
- Removed irrelevant or constant columns: EmployeeNumber, EmployeeCount, Over18, StandardHours.
- Split dataset into **features (X)** and **target (y)**.
- Performed a **70/30 train-test split** using train_test_split.

3. Models Implemented

Logistic Regression

Chosen for its simplicity and interpretability, especially suitable for binary classification like attrition.

Random Forest Classifier

Used for better accuracy and feature importance understanding. It handles non-linear relationships well and avoids overfitting.

Evaluation Metrics Used:

- Confusion Matrix
- Precision, Recall, F1-Score from classification_report

4. SHAP Explanation

To understand the model's decisions, **SHAP (Shapley Additive explanations)** was used on the Random Forest model.

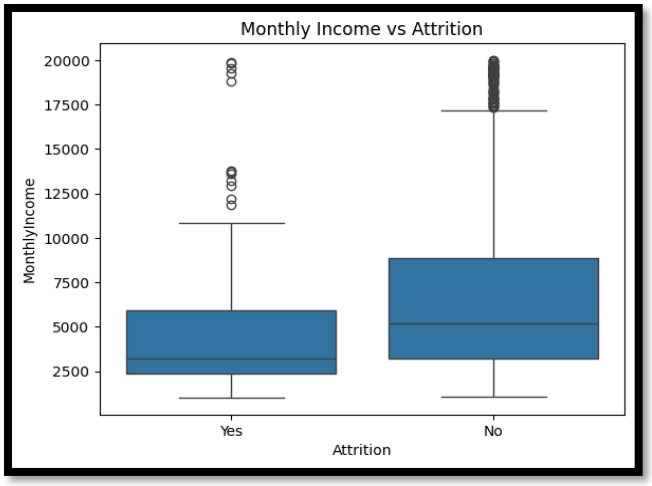
Top Features Identified:

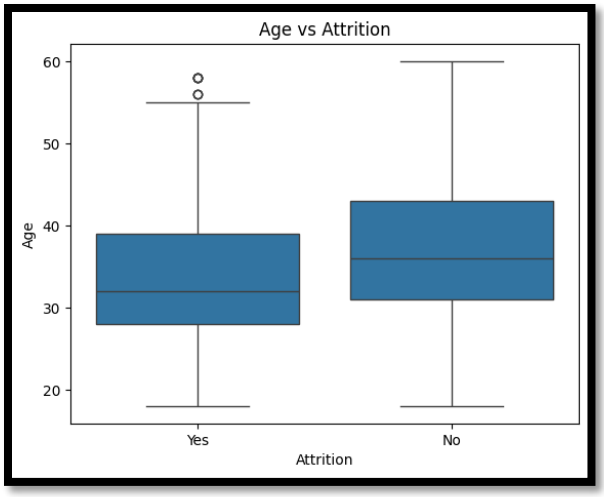
- OverTime
- Age
- JobLevel
- MonthlyIncome
- JobRole

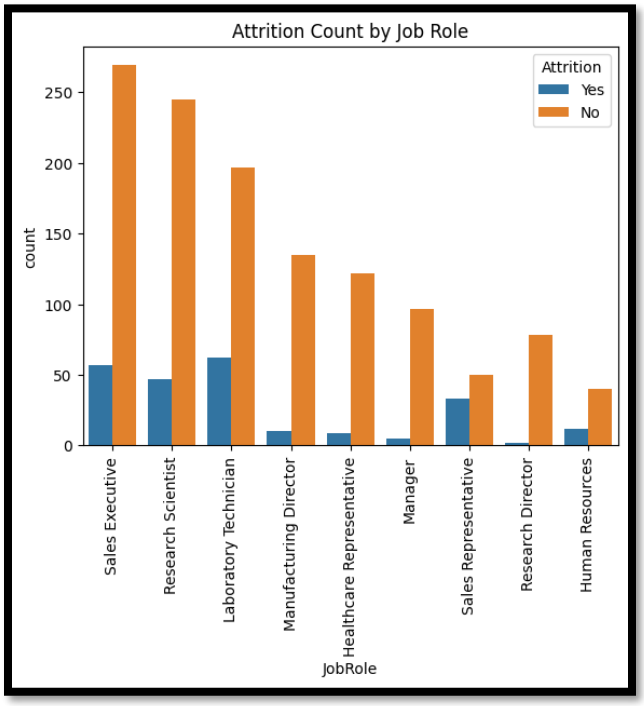
These features had the most impact on whether an employee stayed or left.

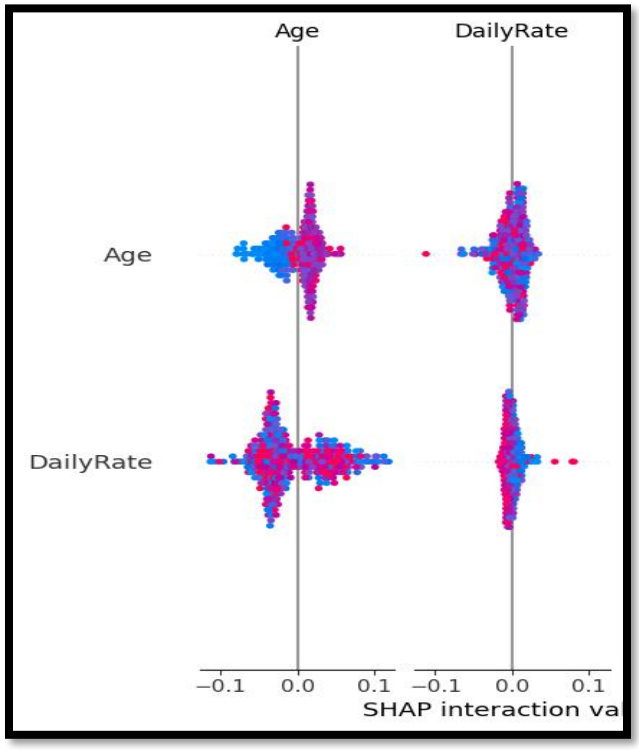
5. Key Insights & Visualizations

- **Overtime** was a major driver of attrition.
- **Low monthly income** and **low job satisfaction** correlated with higher attrition.
- Visualizations (boxplots, countplots) confirmed these patterns.
- SHAP summary plot highlighted the most impactful features.









6. Challenges & Solutions

Challenge	Solution
Class imbalance in Attrition variable	Used model evaluation metrics beyond accuracy

Too many categorical variables	Applied one-hot encoding and dropped multicollinearity
Understanding model decisions	Used SHAP to visualize feature impact

7. Final Recommendations

- Limit excessive overtime to reduce employee burnout
- Improve compensation packages, especially for entry-level roles
- Invest in job satisfaction programs and growth paths
- Monitor high-risk groups (e.g., younger employees, certain job roles)

Conclusion:

This project successfully used machine learning to predict employee attrition and provided data-driven insights to support HR in developing retention strategies.