Employee Attrition Analysis & Prediction

Introduction

Employee attrition poses a major challenge for organizations as it directly impacts productivity, recruitment costs, and operational efficiency. High turnover rates often indicate underlying issues such as job dissatisfaction, inadequate compensation, or limited career progression opportunities.
The purpose of this project is to analyze historical HR data to understand the key factors contributing to attrition and develop predictive models that can help HR teams take proactive measures. By leveraging data analytics and machine learning techniques, this study aims to provide actionable insights and improve employee retention strategies.

Abstract

This project focuses on predicting employee attrition using an HR dataset of approximately 1,470 employees. The analysis involves three key phases:

1. **Exploratory Data Analysis (EDA)** to identify trends and correlations.
2. **Predictive Modeling** using Logistic Regression and Decision Tree algorithms.
3. **Model Explainability** using SHAP values to interpret feature importance.

The final outcome includes a Power BI dashboard for visualization, a Python-based ML pipeline for prediction, and actionable recommendations to reduce attrition rates.

Tools & Technologies Used

- **Python (Pandas, NumPy)** – Data cleaning, preprocessing, and modeling.
- **Visualization**: Matplotlib, Seaborn for EDA; Power BI for interactive dashboards.
- **Machine Learning**: scikit-learn (Logistic Regression, Decision Tree).
- **Model Explainability**: SHAP (SHapley Additive Explanations).

Steps Involved in Building the Project

1. Data Collection and Understanding

- Dataset: *Employee Attrition_dataset.csv* (1,470 records, multiple features including demographics, job role, salary, and performance).
- Target variable: Attrition (Yes/No).

2. Data Preprocessing

- Removed duplicates and handled missing values.
- Applied Label Encoding for categorical variables like Gender, JobRole, and Department.
- Standardized numerical features for Logistic Regression using **StandardScaler**.

- Created a derived feature: **SalaryBand** using quantile-based binning.

3. Exploratory Data Analysis (EDA)

- **Attrition by Department**: Higher attrition observed in Sales and HR.
- **Attrition by Salary Band**: Employees in the lowest salary band show the highest attrition.
- **Promotion Impact**: Longer gaps since last promotion correlate strongly with leaving.

4. Model Development

- **Logistic Regression**:
    - Accuracy: **87.75%**
    - F1-score for Attrition class: **0.53**
- **Decision Tree**:
    - Accuracy: **84.35%**
    - F1-score for Attrition class: **0.28**

5. Explainability with SHAP

- SHAP analysis revealed the top contributing factors:
    - **Monthly Income**, **Years Since Last Promotion**, **Job Level**, and **Age**.

Findings & Insights

- Low salary and lack of promotions significantly increase attrition probability.
- Employees with short tenure and younger age groups are more likely to leave.
- Logistic Regression offers better overall performance and interpretability compared to Decision Trees.

Conclusion

Logistic Regression provided higher accuracy and stability, making it a suitable choice for predicting employee attrition. However, recall for the attrition class is limited, indicating a need for class balancing or threshold adjustments.