Employee Attrition Prediction and Analysis

Milestone 1: Data Collection, Exploration, and Preprocessing

Dataset: Employee_Attrition.csv Rows: ~1,470 employees | Columns: 35

Overview

The objective of Milestone 1 is to prepare the employee dataset for analysis and modeling. This includes collecting data, exploring its structure, and performing preprocessing tasks to ensure readiness for subsequent stages of the project.

---

1. Data Collection

- **Source:** The dataset used for this project contains information on employee demographics, job roles, tenure, performance ratings, salary, and other factors influencing attrition.
- **Number of Records:** 1,470 employees.
- **Number of Features:** 35, comprising both numerical and categorical data.

---

2. Exploratory Data Analysis (EDA)

2.1 Dataset Summary

- **Numerical Features:**
    - Key examples include Age, MonthlyIncome, and YearsAtCompany.
    - Statistical insights:
        - Age: Mean = 36.92, Min = 18, Max = 60
        - MonthlyIncome: Mean = 6,474.98, Min = 1,009, Max = 19,999
        - YearsAtCompany: Mean = 7.01, Min = 0, Max = 40
- **Categorical Features:**
    - Examples: Department, Gender, JobRole, Attrition
    - Attrition breakdown:
        - 1,233 employees stayed (83.9%)

        - 237 employees left (16.1%)

2.2 Missing Values and Duplicates

- No missing values were identified in the dataset.
- No duplicate rows were detected.

2.3 Correlation Analysis

- **Key relationships:**
    - MonthlyIncome and JobLevel: Strong positive correlation (0.95)
    - YearsWithCurrManager and YearsAtCompany: Strong positive correlation (0.77)
    - TotalWorkingYears and JobLevel: Strong positive correlation (0.78)

2.4 Key Patterns and Insights

- Employees in the Sales department had higher attrition rates compared to others.
- Employees with lower monthly incomes and shorter tenures showed a greater likelihood of attrition.
- Strong correlation between job level and salary highlights hierarchical pay structures.

---

3. Preprocessing and Feature Engineering

3.1 Handling Missing Values and Outliers

- No missing data required imputation or removal.
- Outlier detection flagged a small number of extreme values in YearsAtCompany and MonthlyIncome but no immediate action was taken due to their plausibility.

3.2 Feature Engineering

- Encoded categorical variables such as:
    - Gender (e.g., Male = 1, Female = 0)
    - Attrition (e.g., Yes = 1, No = 0)
- Standardized numerical features like MonthlyIncome to ensure consistent scaling.

3.3 Cleaned Dataset

- Delivered a fully cleaned and preprocessed dataset, ready for advanced analysis and model building in Milestone 2.
