EMPLOYEE ATTRITION PREDICTION

By
Shirin Fathima A

Comprehensive Report on Employee Attrition Analysis

Introduction

Employee attrition is a critical challenge faced by organizations, impacting operational efficiency, productivity, and financial stability. This project aims to analyze the factors contributing to employee attrition and develop a predictive model to identify employees at risk of leaving the organization. Using a dataset from a human resource management system, we perform a structured analysis involving data preprocessing, exploratory data analysis (EDA), and machine learning modeling. The insights derived from this project can help organizations design targeted retention strategies and foster a better work environment.

Dataset Overview

The dataset used in this project contains a comprehensive set of features that include:

- **Demographics**: Age, Gender, Marital Status.
- **Work Environment**: Business Travel, Department, Job Role.
- **Performance Metrics**: Monthly Income, Performance Rating.
- **Other Factors**: Overtime, Distance from Home, Work-Life Balance.

The dataset has the following characteristics:

- **Total Rows**: 1,470.
- **Total Columns**: 35.

Initial Exploration

- The dataset was loaded using Pandas and inspected for column names, data types, and summary statistics.
- Features such as `EmployeeCount` and `StandardHours`, which provided no variance or relevant information, were removed.
- No significant missing values were present in the dataset, ensuring data completeness.

Data Cleaning and Preprocessing

1. **Handling Irrelevant Columns**:
    - Columns like `EmployeeCount`, `StandardHours`, and `EmployeeNumber` were removed as they did not contribute meaningful information to the analysis.
2. **Encoding Categorical Variables**:

- Features such as `Attrition`, `OverTime`, and `Gender` were mapped to numerical values (e.g., Yes/No converted to 1/0).
- One-hot encoding was applied to multi-class categorical variables like `Department`, `EducationField`, and `JobRole`.

3. **Balancing the Dataset**:
    - The target variable, `Attrition`, was imbalanced, with more "No" (employees staying) than "Yes" (employees leaving).
    - Random oversampling was applied to create a balanced dataset, ensuring the model does not favor the majority class.
4. **Feature Scaling**:
    - StandardScaler was used to standardize numerical features, normalizing data to have a mean of 0 and a standard deviation of 1. This step ensures that all features contribute equally to the machine learning model.

Exploratory Data Analysis (EDA)

EDA was conducted to uncover relationships between features and attrition. Key observations include:

1. **Attrition Distribution**:
    - The attrition rate was approximately 16%, indicating the majority of employees stayed.
2. **Impact of Overtime**:
    - Employees working overtime showed a significantly higher attrition rate.
3. **Business Travel**:
    - Frequent travel correlated with increased attrition rates.
4. **Income Levels**:
    - Lower-income employees had a higher likelihood of leaving compared to higher-income employees.

Visualizations such as count plots, box plots, and scatter plots were generated using Seaborn to illustrate these trends.

Machine Learning Model

Model Selection

Gradient Boosting Classifier was chosen for its ability to handle imbalanced datasets and deliver high accuracy with minimal overfitting.

Steps Involved

1. **Data Splitting**:
    - The dataset was split into training (80%) and testing (20%) subsets.
2. **Model Training**:
    - The Gradient Boosting Classifier was trained using 200 estimators.
    - Training accuracy: **98.6%**
    - Testing accuracy: **87.4%**
3. **Evaluation**:
    - Confusion Matrix:
        - True Positives: Employees correctly identified as leaving.
        - True Negatives: Employees correctly identified as staying.
        - False Positives and False Negatives were minimal, indicating good performance.
    - Classification Report:
        - Precision: **0.85**
        - Recall: **0.88**
        - F1-Score: **0.86**

Key Insights and Results

1. **Overtime**:
    - A strong predictor of attrition. Employees working overtime are more likely to leave.
2. **Low Income**:
    - Financial dissatisfaction plays a significant role in attrition.
3. **Frequent Travel**:
    - Employees frequently traveling for work are at higher risk.
4. **Work-Life Balance**:
    - Poor work-life balance contributes significantly to employee dissatisfaction.

Single Prediction Demonstration

To showcase the model's applicability, a prediction was made for a single employee based on their features. The model outputs:

- **Satisfied Employee**: Employees unlikely to leave.

- **Unsatisfied Employee**: Employees at risk of leaving.

Example Output:

> "Employee is satisfied with the company" or "Employee is not satisfied with the company."

Conclusion

This project demonstrates the power of data-driven approaches in addressing employee attrition. By leveraging machine learning, organizations can proactively identify employees at risk of leaving and implement strategies to improve retention. The Gradient Boosting Classifier proved to be an effective tool for this task, achieving high accuracy and providing actionable insights. Future improvements could include testing additional models, incorporating advanced feature selection methods, and exploring external factors like market conditions.

Future Scope

1. **Model Optimization**:
    - Hyper parameter tuning to improve model performance.
2. **Feature Engineering**:
    - Adding interaction terms or external datasets to enhance predictive power.
3. **Real-Time Prediction**:
    - Integrating the model into HR systems for continuous monitoring and real-time decision-making.