Report: HR Analytics: Employee Attrition Prediction

By - Anjali Gupta

Introduction

Employee attrition is a critical challenge faced by organizations as it directly impacts productivity, employee morale, and operational costs. High attrition rates lead to increased hiring and training expenses, loss of experienced talent, and disruption in team performance. Understanding the factors that influence employees to leave an organization is essential for developing effective retention strategies.

This project focuses on analyzing employee data to identify key drivers of attrition and predict the likelihood of employees leaving the organization. By applying data analytics and machine learning techniques, the project aims to provide actionable insights that can help HR departments reduce employee turnover and improve workforce stability.

Abstract

The objective of this project is to analyze employee attrition patterns and build a predictive model that identifies employees at risk of leaving the organization. Using an HR attrition dataset, exploratory data analysis was performed to understand trends related to department, income, overtime, tenure, and job-related factors. A Logistic Regression classification model was developed to predict attrition, and class imbalance was handled using the SMOTE technique. Model performance was evaluated using accuracy, confusion matrix, and ROC–AUC metrics. Additionally, SHAP value analysis was applied to interpret model predictions and identify the most influential features contributing to attrition. The insights derived from this analysis can support data-driven HR decision-making and employee retention strategies.

Tools Used

- **Python** – Core programming language used for analysis and modeling
- **Pandas & NumPy** – Data manipulation and numerical operations
- **Matplotlib & Seaborn** – Data visualization and exploratory analysis
- **Scikit-learn** – Machine learning model building and evaluation
- **Imbalanced-learn (SMOTE)** – Handling class imbalance in attrition data
- **SHAP** – Model explainability and feature importance analysis
- **Google Colab** – Cloud-based environment for execution

Steps Involved in Building the Project

1. **Data Loading and Preprocessing**
   The HR attrition dataset was loaded and inspected to understand its structure, data types, and completeness. The target variable *Attrition* was converted into numerical

format, and non-informative or constant columns were removed. Categorical variables were encoded, and numerical features were standardized to prepare the data for modeling.

2. **Exploratory Data Analysis (EDA)**
   EDA was performed to identify patterns and trends related to employee attrition. Visual analysis revealed that attrition was higher in Sales and Research & Development departments. Employees with lower monthly income, fewer years at the company, and those working overtime showed higher attrition rates. A correlation heatmap further highlighted relationships between attrition, income, job level, and tenure-related features.
3. **Model Building**
   A Logistic Regression model was trained to predict employee attrition. The dataset was split into training and testing sets to evaluate model performance on unseen data.
4. **Handling Class Imbalance**
   Since the dataset was imbalanced with fewer attrition cases, SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the training data. The Logistic Regression model was retrained on the balanced dataset to improve prediction of the minority class.
5. **Model Evaluation**
   Model performance was assessed using accuracy, classification report, confusion matrix, and ROC–AUC score. The SMOTE-based model showed improved recall for attrition cases, indicating better identification of employees at risk.
6. **Model Explainability using SHAP**
   SHAP value analysis was conducted to interpret the model's predictions. The SHAP summary and bar plots identified key factors influencing attrition, such as overtime, monthly income, job level, years at company, and work-life balance.

Conclusion

This project successfully analyzed employee attrition using data analytics and machine learning techniques. The results highlight that compensation, overtime, tenure, and career progression play significant roles in employee turnover. Employees working overtime, earning lower income, and having fewer years at the company are more likely to leave the organization.

The predictive model, combined with SHAP explainability, provides transparent and actionable insights for HR teams. Based on the analysis, organizations can reduce attrition by improving compensation structures, monitoring overtime workloads, supporting early-tenure employees, and creating clear career growth opportunities. Overall, this project demonstrates how data-driven HR analytics can support informed decision-making and enhance employee retention strategies.