Employee Attrition Analysis Report

Introduction:

This report analyzes employee attrition using the IBM HR Employee Attrition dataset. The analysis aims to identify factors contributing to attrition and build predictive models to forecast employee turnover.

Data Exploration and Preprocessing:

- The dataset contains information about employee demographics, job-related factors, and attrition status.

- Initial exploration involved examining the dataset's structure, descriptive statistics, and distributions of key variables.

- Missing values and duplicate entries were checked, with no significant issues found.

- Categorical variables were encoded using Label Encoding to prepare them for modeling.

- The dataset was split into training and testing sets (80% train, 20% test) for model evaluation.

- Standard scaling was applied to numerical features to ensure consistent model performance.

Exploratory Data Analysis (EDA):

EDA was conducted to visualize relationships between various factors and attrition:

- Department: Attrition rates varied across departments, with Sales and Research & Development having higher rates.

- Age: The distribution of age for employees who left was slightly lower than those who stayed.

- Distance from Home: Employees living farther from home showed a higher likelihood of attrition.

- Monthly Income: Lower monthly income was associated with higher attrition rates.

- Job Satisfaction: Lower job satisfaction levels were linked to increased attrition.

- Work-Life Balance: Poorer work-life balance was associated with higher attrition.

- Overtime: Employees working overtime were more likely to leave.

- Years at Company: Attrition was more pronounced among employees with fewer years at the company.

- Job Role: Certain job roles, such as Sales Representatives and Laboratory Technicians, exhibited higher attrition.

- Performance Rating: Interestingly, there was no strong correlation between performance rating and attrition.

Model Building and Evaluation:

Two models were trained to predict attrition: Logistic Regression and Random Forest.

- Logistic Regression achieved an accuracy of approximately 87%.

- Random Forest outperformed Logistic Regression, achieving an accuracy of around 84%.

Feature Importance (SHAP):

SHAP (Shapley Additive explanations) was used to interpret model predictions and identify important features.

- Key factors influencing attrition, as revealed by SHAP, included overtime, monthly income, job satisfaction, age, and total working years.

- SHAP summary plots and feature importance plots provided insights into the directional impact of each feature on attrition predictions.

Conclusion:

The analysis revealed valuable insights into employee attrition drivers. Factors like overtime, job satisfaction, and income play crucial roles in predicting employee turnover. The Random Forest model demonstrated promising predictive performance.

Recommendations:

Based on the findings, the following recommendations are made:

- Address factors contributing to attrition: Focus on improving work-life balance, increasing job satisfaction, and ensuring competitive compensation.

- Retention strategies: Implement targeted retention programs for high-risk employee groups, such as those in specific departments or job roles.

- Use predictive models: Leverage the developed models to identify employees at risk of leaving and proactively address their concerns.