Analysis Report: Employee Attrition Prediction

Main Objective: The analysis aims to build a predictive classification model to identify employees at risk of attrition. The primary objective is prediction-oriented, offering actionable insights for HR management to proactively retain talent, reduce hiring costs, and maintain organizational stability.

Data Set Description: The chosen dataset is the IBM Employee Attrition dataset, containing 1,470 observations and 35 attributes, including demographics (age, gender, marital status), job-related details (job role, department, years at company), compensation metrics (salary, stock options), and employee satisfaction indicators (job satisfaction, work-life balance). The target variable, "Attrition," indicates whether an employee has left the company (Yes/No).

The goal is to accurately predict employee attrition and uncover insights into primary drivers influencing an employee's decision to leave.

Data Exploration, Cleaning, and Feature Engineering: Initial exploration revealed no missing values. Attributes with low variability or redundancy (e.g., EmployeeNumber, Over18, StandardHours) were removed. Categorical variables were encoded using One-Hot Encoding, while numerical variables underwent standardization.

Feature engineering involved creating derived features like "TotalWorkingYearsGrouped," "IncomeSatisfactionRatio," and "PromotionInLast5Years." The final dataset comprised 30 predictive features after cleaning and engineering.

Classifier Models Summary: Three classifier models were trained using a 70/30 train-test split:

1. **Logistic Regression:** Achieved an accuracy of 88%. Despite high interpretability, this model showed limitations in recall for employees leaving (42%) with an F1-score of 0.53.
2. **Random Forest:** Provided an accuracy of 84%. The recall for attrition cases was notably low (17%), indicating difficulty in accurately identifying employees likely to leave. Top features included "MonthlyIncome," "Age," "TotalWorkingYears," and "YearsAtCompany."
3. **Gradient Boosting Classifier (XGBoost):** Achieved an accuracy of 85%. Although slightly lower accuracy than Logistic Regression, it improved recall (27%) for attrition cases. Important features included "Department_Human Resources," "StockOptionLevel," "EducationField_Human Resources," and "TotalWorkingYears."

Recommended Classifier: The Logistic Regression model is recommended due to its balanced accuracy and interpretability. Although its recall for identifying attrition is limited, its high precision and ease of interpretation make it suitable for HR decision-making, especially when combined with other retention strategies.

Key Findings and Insights:

- Important predictors of employee attrition include:
    - Monthly income
    - Age and tenure-related metrics (total working years, years at the company)
    - Job role and department, particularly Human Resources roles
    - Stock option availability and overtime practices
- Younger employees or those with fewer years at the company have higher attrition risks.
- Non-financial factors such as departmental role, career progression, and overtime contribute significantly to employee turnover.

Suggestions for Next Steps:

- Improve predictive capability for minority class using specialized methods such as SMOTE or class weighting.
- Integrate additional qualitative data, including employee engagement surveys or exit interview insights.
- Explore interactions between key variables, especially departmental characteristics and compensation.
- Regularly review and recalibrate models with updated internal and external market data.