Employee Attrition Analysis Report

1. Dataset Overview

The dataset consists of 74,610 employee records with various features related to demographics, work history, job satisfaction, education, and company attributes.

Numerical features include Age, Years at Company, Monthly Income, Number of Promotions, Distance from Home, and Number of Dependents.

Categorical features include variables like Job Role, Gender, Work-Life Balance, Education Level, and Attrition status.

The attrition classes are nearly balanced with approximately 52.5% employees staying and 47.5% leaving, making the dataset suitable for binary classification modeling.

There are minimal missing values: only ~2.5% in 'Distance from Home' and ~3.2% in 'Company Tenure'.

2. Exploratory Data Analysis (EDA)

Histograms and boxplots reveal that:

- Age is normally distributed with no significant outliers.
- Years at Company is right-skewed, indicating more new employees.
- Monthly Income is bimodal and contains significant high-end outliers.
- Number of Promotions and Number of Dependents are heavily right-skewed.

The correlation heatmap shows a moderate correlation (0.53) between Age and Years at Company, indicating tenure grows with age.

Other numerical variables show weak or no correlation with one another, implying independence across most features.

3. Categorical Feature Distributions

Employee Attrition Analysis Report

Majority of employees are Male (55%) and in Technology roles (26%).

Most employees rate their Work-Life Balance as 'Good' or 'Fair' and Job Satisfaction as 'High'.

The most common education level is a Bachelor's Degree.

Overtime appears common (33%), and most employees are not remote workers.

Only a small fraction (5%) reported access to Leadership Opportunities, indicating a potential driver of attrition.

4. Bivariate Analysis on Attrition

- Attrition is higher among Females and Single employees.

- Poor Work-Life Balance and Low Job Satisfaction significantly correlate with higher attrition.

- Overtime is a major factor: employees working overtime have significantly higher attrition.

- Employees without remote work access or with poor company reputation also show elevated attrition rates.

- Interestingly, lower attrition is observed among Senior-level employees and those with PhDs, possibly due to better roles or stability.

5. Feature Engineering

Features were encoded using one-hot encoding. Validation and training sets were aligned for consistent column structure.

The final training data had 39 features, including encoded versions of categorical variables such as Job Role, Recognition, Reputation, etc.

Target variable 'Attrition' was binarized where 1 = Stayed and 0 = Left.

Feature standardization ensured zero-mean and unit variance for numerical columns.

6. Feature Selection & Multicollinearity

Employee Attrition Analysis Report

RFE (Recursive Feature Elimination) selected the top 15 features contributing most to the model.

Important predictors include Distance from Home, Overtime, Gender, Education Level, and Job Satisfaction.

VIF analysis confirmed absence of multicollinearity (all VIF < 1.3), indicating predictors are independent.

7. Model Summary (Logistic Regression)

A logistic regression model was trained using the selected features.

Pseudo R-squared: 0.278 indicates moderate explanatory power.

Key predictors with positive coefficients: Distance from Home, Poor Work-Life Balance, Low Recognition, and Single Marital Status.

Negative coefficients: Senior Job Level, Remote Work, and Education Level (PhD).

Most coefficients are statistically significant ($p < 0.05$), suggesting strong associations with attrition likelihood.

8. Model Evaluation

ROC AUC score of 0.83 indicates strong model performance.

Precision-Recall AP score of 0.817 further validates its ability to balance false positives and negatives.

Sensitivity (Recall) and Specificity were visualized against different probability thresholds.

An optimal balance around a threshold of ~0.5 provides 73% accuracy on training set, with sensitivity and specificity around 70-75%.

9. Conclusion

The logistic regression model effectively captures attrition patterns with high interpretability and solid performance.

Employee Attrition Analysis Report

Key takeaways:

- Employees working overtime and lacking remote work privileges are more likely to leave.

- Improving work-life balance, recognition programs, and career growth opportunities may reduce attrition.

- Demographics such as gender and marital status also influence employee stability and retention.

Authored By

Soumyajit Bera