Predicting Employee Retention

1. Introduction

Problem Statement

Employee retention is a critical issue for organizations, influencing not only productivity but also team cohesion, corporate culture, and the financial bottom line. High attrition rates can lead to increased recruitment and training costs, disruption of ongoing projects, and loss of institutional knowledge. By accurately predicting employee retention, organizations can proactively address the root causes of turnover and implement targeted strategies to maintain a stable and engaged workforce.

Objective

The primary goal of this project is to develop a Logistic Regression model capable of predicting employee retention based on diverse employee-related variables. These include demographic information, job satisfaction scores, performance ratings, compensation details, and tenure. The intention is to equip the HR department with data-driven insights that will help foster a more supportive, inclusive, and engaging work environment conducive to employee longevity.

2. Methodology

Data Collection

The dataset comprises 74,610 employee records with 24 distinct features. These features span multiple domains—ranging from personal attributes (age, gender) to professional metrics (monthly income, job level, performance ratings), offering a holistic view of the workforce.

Data Preprocessing

- Handling Missing Values: Rows with missing or inconsistent values were removed after careful evaluation of their impact.
- Redundant Columns: Features such as 'Employee ID', 'Overtime', and 'Company Reputation' were excluded.
- Outlier Analysis: Extreme values in 'Number of Dependents' and 'Monthly Income' exceeding the 99th percentile were eliminated.
- Train-Test Split: The dataset was divided into training and testing subsets using a 70:30 ratio, ensuring balanced distribution.

3. Techniques Used

Machine Learning Models

A Logistic Regression model was implemented for its interpretability and suitability for binary classification.

Evaluation Metrics

- Precision: Reliability of positive predictions (i.e., predicted to stay).
- Specificity: Model's ability to identify employees unlikely to stay.

- Sensitivity (Recall): Model's ability to identify those who will stay.
- Accuracy: Overall correctness of classification.
- ROC Curve: Trade-off between true positive rate and false positive rate.

---

4. Analysis

Univariate Analysis

- Experience: Skew toward 0–3 years highlights need for improved onboarding.
- Work-Life Balance: Variability indicates room for policy improvement.
- Age: Majority in the 25–35 range; mid-career development is key.
- Monthly Income: Bell-shaped curve suggests scope for compensation benchmarking.
- Job Satisfaction: Medium to high satisfaction; low scorers need attention.

Bivariate Analysis

- Experience vs. Attrition: High turnover among less experienced.
- Age vs. Attrition: Younger employees need growth opportunities.
- Monthly Income vs. Attrition: Salary dissatisfaction is a key factor.
- Job Satisfaction vs. Attrition: Emotional engagement is crucial.
- Work-Life Balance vs. Attrition: Flexible hours and wellness matter.

---

5. Outlier Analysis

- Number of Dependents: Extreme values removed to avoid skew.
- Monthly Income: Top earners filtered to reflect representative data.

---

6. Redundant Column Explanation

- Employee ID: No predictive value.
- Overtime: Redundant with other workload variables.
- Company Size: Covered by department and level.

- Company Tenure (In Months): Correlated with 'Years at Company'.
- Remote Work: Minimal variation.
- Job Role: High cardinality and overlap.
- Company Reputation: Subjective and inconsistently recorded.

---

7. Key Insights

Feature Selection

Recursive Feature Elimination (RFE) highlighted 'Job Satisfaction', 'Monthly Income', 'Age', and 'Work-Life Balance' as strong predictors.

Model Building

Logistic Regression trained with selected features showed statistically significant coefficients.

Model Evaluation

- Training Accuracy: 0.85
- Test Accuracy: 0.83

Confusion Matrix Analysis:

- TP = Correctly predicted retained employees
- TN = Correctly predicted non-retained employees
- FP = Incorrectly predicted retained employees
- FN = Incorrectly predicted non-retained employees
- Sensitivity: 0.73
- Specificity: 0.70
- Precision: 0.72
- Recall: 0.73

The model demonstrated a balanced capacity to detect both retention and attrition.

---

8. Conclusion

Summary

The Logistic Regression model successfully predicted employee retention by leveraging simplified yet powerful features. Key factors include job satisfaction, income, experience, and work-life balance.

Future Work

- Incorporate advanced models like Random Forest and Neural Networks.
- Analyze temporal trends using time-series data.
- Enrich qualitative insights with sentiment analysis.
- Employ cross-validation for better generalization.
