Predicting Employee Retention

1. Introduction

Problem Statement

Employee retention is a critical issue for organizations, impacting productivity, morale, and financial performance. Predicting employee retention can help organizations proactively address factors leading to attrition and implement strategies to retain valuable employees.

Objective

The objective of this project is to develop a Logistic Regression model to predict employee retention based on various factors such as demographic details, job satisfaction scores, performance metrics, and tenure. The goal is to provide actionable insights to the HR department to strengthen retention strategies and create a supportive work environment.

2. Methodology

Data Collection

The dataset consists of 24 columns and 74,610 rows, including demographic details, job satisfaction scores, performance metrics, and tenure.

Data Preprocessing

- **Handling Missing Values**: Missing values were identified and handled by dropping rows with missing data.
- **Redundant Columns**: Columns such as 'Employee ID', 'Overtime', 'Company Size', 'Company Tenure (In Months)', 'Remote Work', 'Job Role', and 'Company Reputation' were dropped as they were deemed redundant.
- **Outlier Analysis**: Outliers in numerical columns were identified using box plots and handled by removing rows exceeding the 99th percentile for 'Number of Dependents' and 'Monthly Income'.

Train-Test Split

The data was split into 70% training data and 30% test data.

3. Techniques Used

Machine Learning Models

A Logistic Regression model was used to predict employee retention.

Evaluation Metrics

- **Precision**: Measures the proportion of true positives among the predicted positives.
- **Specificity**: Measures the proportion of true negatives among the actual negatives.
- **Sensitivity (Recall)**: Measures the proportion of true positives among the actual positives.
- **Accuracy**: Measures the overall correctness of the model.
- **ROC (Receiver Operating Characteristic)**: Evaluates the trade-off between sensitivity and specificity.

4. Analysis

Univariate Analysis

Univariate analysis involves examining each variable individually to understand its distribution and identify any patterns or anomalies.

1. Experience

The distribution of employee experience shows a significant count with 0-3 years and a lesser count above 20 years. This indicates that most employees have relatively low experience, which could be a factor in attrition.

2. Work-Life Balance

The work-life balance variable is categorized into Poor, Below Average, Good, and Excellent. The analysis shows a higher number of counts in specific categories, indicating varying levels of satisfaction among employees.

3. Age

The age distribution of employees ranges from 18 to 60 years. Most employees fall within the 25-35 age range, which could be a critical factor in retention strategies.

4. Monthly Income

Monthly income varies significantly among employees, with most earning between $3000 and $7000. Understanding income distribution helps in identifying financial satisfaction and its impact on retention.

5. Job Satisfaction

Job satisfaction is categorized into Very Low, Low, Medium, and High. The analysis reveals that most employees report Medium to High satisfaction, which is crucial for retention.

Bivariate Analysis

Bivariate analysis examines the relationship between two variables to understand how they interact and influence each other.

1. Experience vs Attrition

The analysis shows that employees with 0-3 years of experience are more likely to leave, while those with higher experience tend to stay. This indicates that newer employees may need more support and engagement to improve retention.

2. Age vs Attrition

Younger employees (18-25 years) have higher attrition rates compared to older employees (35-60 years). This suggests that younger employees may be more prone to job changes and require targeted retention strategies.

3. Monthly Income vs Attrition

Employees with lower monthly income are more likely to leave, indicating that financial dissatisfaction could be a significant factor in attrition. Higher income levels correlate with lower attrition rates.

4. Job Satisfaction vs Attrition

Employees with Very Low to Low job satisfaction have higher attrition rates, while those with Medium to High satisfaction are more likely to stay. This underscores the importance of job satisfaction in retention.

5. Work-Life Balance vs Attrition

Employees reporting Poor to Below Average work-life balance have higher attrition rates. Improving work-life balance could be a key strategy in reducing turnover.

5. Outlier Analysis

Outliers can significantly impact the results of data analysis and model performance. Identifying and handling outliers ensures the accuracy and reliability of the model.

1. Number of Dependents

Outliers in the 'Number of Dependents' feature were identified using box plots. Rows where the number of dependents exceeded the 99th percentile were removed to ensure a more accurate analysis.

2. Monthly Income

Similarly, outliers in the 'Monthly Income' feature were identified and handled by removing rows exceeding the 99th percentile. This helps in maintaining the integrity of the data.

6. Redundant Column Explanation

Redundant columns were identified and dropped to streamline the dataset and improve model performance.

1. Employee ID

The 'Employee ID' column was dropped as it is a unique identifier and does not contribute to the prediction of retention.

2. Overtime

The 'Overtime' column was dropped due to its redundancy in the context of predicting retention.

3. Company Size

The 'Company Size' column was dropped as it does not directly influence individual employee retention.

4. Company Tenure (In Months)

The 'Company Tenure (In Months)' column was dropped as it is redundant with the 'Years at Company' column.

5. Remote Work

The 'Remote Work' column was dropped as it does not significantly impact retention prediction.

6. Job Role

The 'Job Role' column was dropped to simplify the dataset and focus on more relevant features.

7. Company Reputation

The 'Company Reputation' column was dropped due to its redundancy in the context of predicting retention.

7. Key Insights

Feature Selection

Recursive Feature Elimination (RFE) was used to select the top 15 features for the model. This method helps in identifying the most influential features for predicting retention.

Model Building

A Logistic Regression model was built using the selected features. The model's coefficients and p-values were evaluated to ensure statistical significance.

Model Evaluation

- **Accuracy**: The model achieved an accuracy of 0.85 on the training set and 0.83 on the test set.
- **Confusion Matrix**: The confusion matrix was used to calculate true positives, true negatives, false positives, and false negatives.
    - **True Positive (TP)**: Number of correctly predicted positive cases.
    - **True Negative (TN)**: Number of correctly predicted negative cases.
    - **False Positive (FP)**: Number of incorrectly predicted positive cases.
    - **False Negative (FN)**: Number of incorrectly predicted negative cases.
- **Sensitivity and Specificity**: The model's sensitivity and specificity were calculated to evaluate its performance. Sensitivity = 0.73, Specificity = 0.7
- **Precision and Recall**: Precision and recall values were calculated to assess the model's effectiveness. Precision = 0.72, Recall = 0.73

8. Conclusion

Summary

The Logistic Regression model successfully predicted employee retention with high accuracy. Key factors influencing retention were identified, providing valuable insights for the HR department.

Future Work

Further research could explore additional features and advanced machine learning models to improve prediction accuracy.