Lead Scoring Case Study

Tejaswini V, Koushik Bhakat, Tejus B

Executive summary

- Traditionally, the organization has addressed turnover reactively, relying on exit interviews and historical trends. To proactively identify employees at risk of leaving—and the factors driving attrition—this project develops a logistic-regression model using demographic details, job-satisfaction scores, performance metrics and tenure. By predicting which employees are likely to stay, HR can target retention strategies, bolster engagement, and reduce hiring costs and productivity loss.

Table Of Contents

Data Understanding

Data load & structure

- The original dataset (after initial import) contained employee records with demographic, job-related and performance-related columns.
- Summary statistics (mean, median, ranges) were computed for numerical features (e.g., Age, MonthlyIncome, TotalWorkingYears) and value counts for categoricals (e.g., Gender, Department).

Initial observations

- Attrition rate was approximately 16–18%.
- Numerical variables like MonthlyIncome and YearsAtCompany exhibited right-skewed distributions.
- Several categorical features (e.g., OverTime, WorkLifeBalance) had actionable levels.

Data Cleaning

Handling missing values

- Rows with any missing entries were dropped, reducing the dataset to **70,635 employees** (from ~73,000).

Redundant categorical values

- Standardized categories (e.g., harmonized “Yes”/“YES” in OverTime, unified department names).

Dropping redundant columns

- Removed identifier columns (EmployeeNumber, EmployeeCount) and constant fields that do not contribute to prediction.

Train–Validation Split

The cleaned data was split **70:30** (train vs. validation) using `train_test_split(random_state=42)`.

Training set: ~49,445 rows

Validation set: ~21,190 rows

EDA on Training Data

Univariate analysis

- **Age**: mean ~36 years; slight right skew.
- **MonthlyIncome**: median ~6,000; a few high earners.
- **Attrition**: ~17% “Yes,” indicating moderate class imbalance.

Correlation analysis

- Moderate positive correlation between **TotalWorkingYears** and **YearsAtCompany**.
- Weak correlation between **Education** level and attrition.

EDA on Training Data

Class balance

- Positive (attrition = 1): ~17%
- Negative (attrition = 0): ~83%

Bivariate analysis

- Higher attrition among employees who work overtime.
- Lower job satisfaction (“Low” or “Fair”) and poor work-life balance associate with higher attrition rates.
- Departments like Sales and Research & Development showed elevated turnover compared to HR

Feature Engineering

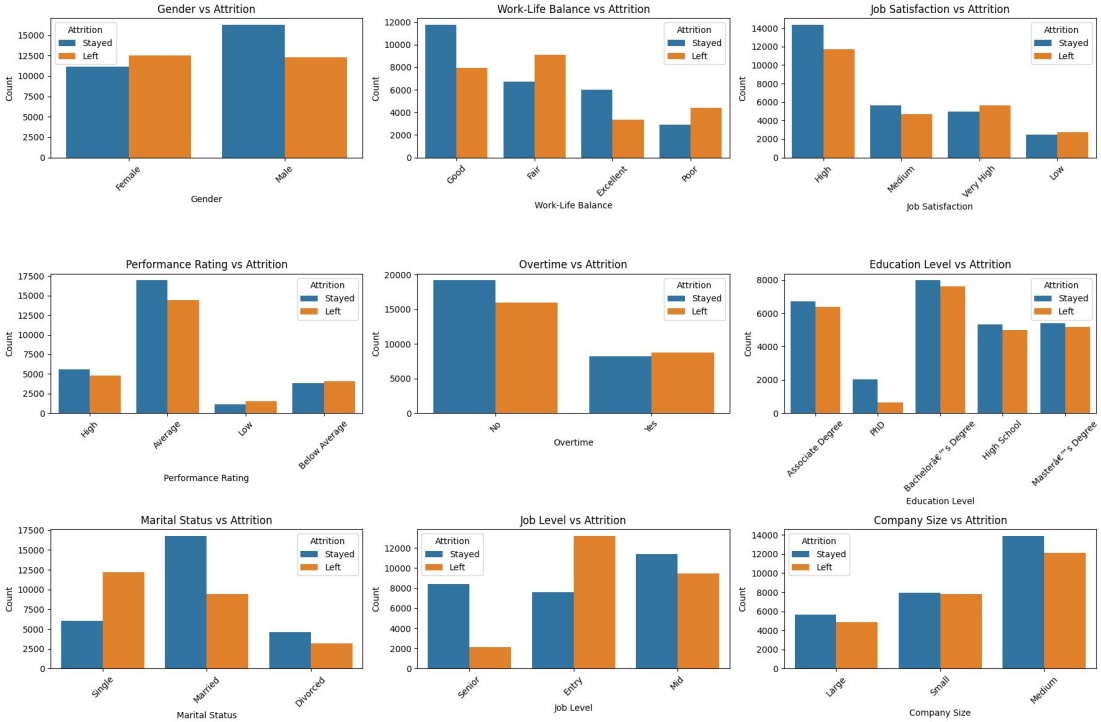
Dummy variable creation

- Converted categorical features (Gender, Department, EducationField, OverTime, WorkLifeBalance levels, etc.) into one-hot encoded columns.

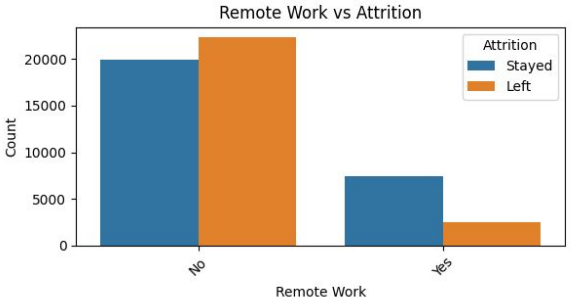
Feature scaling

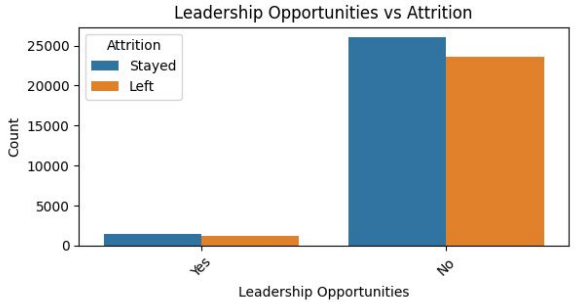
- Applied `StandardScaler` to numerical variables (Age, MonthlyIncome, DistanceFromHome, etc.) to ensure comparability

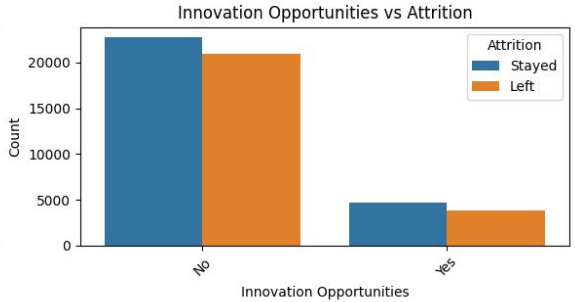
Visualisations

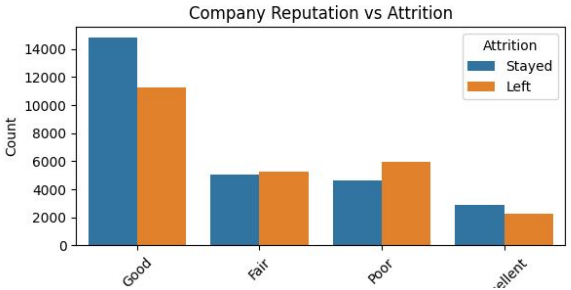


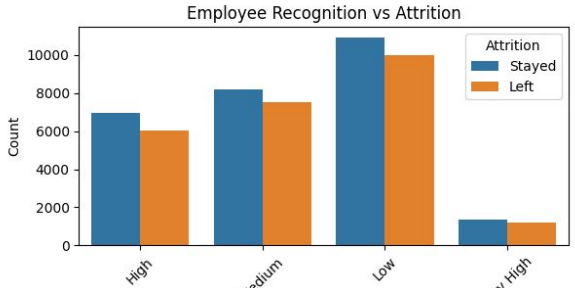
Visualisations











Model Building

Feature selection

- Employed Recursive Feature Elimination (RFE) with a logistic-regression estimator to select the **top 15 features**.
- **Notable selected features:** OverTime_Yes; WorkLifeBalance_Fair/Poor; MonthlyIncome; Age; TotalWorkingYears; JobSatisfaction levels; CompanyReputation_Poor/Fair.

Logistic regression

- Trained `LogisticRegression(random_state=42, max_iter=1000)` on the reduced feature set.

Optimal cutoff

- Evaluated thresholds from 0.1 to 0.9 on the training set; selected **0.50** as the balance point maximizing sensitivity while retaining acceptable specificity.
- Final training accuracy at cutoff 0.50: **71.86%**.

Prediction and Model Evaluation

Metric	Validation Set
Accuracy	73.99%
Confusion Matrix	TN = 8,827
	FN = 2,988
Sensitivity (Recall)	76.82%
Specificity	66.59%
Precision	71.38%
Recall	76.82%

Key Insights & Recommendations

Overtime and **poor work-life balance** are the strongest predictors of attrition—target policies to manage workload and flexible scheduling.

Employees with **lower job-satisfaction** or negative perceptions of company reputation exhibit higher churn—prioritize engagement surveys and tailored career-development plans.

Longer tenure and **higher income** correlate with retention—consider structured progression paths and competitive compensation to encourage loyalty.