Employee Attrition Analysis and Prediction Report

---

1. Problem Statement

Employee attrition is a critical challenge for organizations, leading to increased recruitment costs, loss of institutional knowledge, and reduced team morale. This project aims to:

- **Predict attrition**: Build a model to classify employees as "likely to leave" or "likely to stay."
- **Identify drivers**: Uncover key factors influencing attrition to inform retention strategies.
- **Provide actionable insights**: Recommend interventions to reduce turnover and improve employee satisfaction.

---

2. Methodology

The analysis followed a structured workflow:

1. **Data Understanding**:
   - Explored dataset structure, variables, and missing values.
   - Analyzed descriptive statistics (e.g., mean age, income distribution).
2. **Data Cleaning**:
   - Handled missing values via median (numerical) and mode (categorical) imputation.
   - Fixed encoding errors in categorical columns (e.g., "Bachelor's Degree").
   - Dropped non-informative columns (e.g., Employee ID).
3. **Train-Validation Split**:
   - Split data into **70% training** (34,611 observations) and **30% validation** (14,833 observations) sets.
   - Ensured class balance preservation (Stayed = 83%, Left = 17%).
4. **Exploratory Data Analysis (EDA)**:
   - Performed univariate, bivariate, and correlation analysis.
   - Visualized trends and relationships (see **Section 5**).
5. **Feature Engineering**:
   - Created dummy variables for categorical features (e.g., Gender, Job Role).
   - Scaled numerical features using standardization.
6. **Model Building**:
   - Trained a **logistic regression model** (GLM with binomial family and logit link).
   - Evaluated performance using accuracy, sensitivity, specificity, and precision.
7. **Validation**:
   - Tested model generalizability on unseen data.

    - Analyzed feature importance via coefficients and odds ratios.

---

3. Techniques Used

Data Preprocessing

- **Missing Values**: Median imputation for numerical features (e.g., Monthly Income), mode imputation for categorical features (e.g., Education Level).
- **Categorical Encoding**: One-hot encoding for variables like Job Satisfaction and Work-Life Balance.
- **Feature Scaling**: Standardized numerical features (e.g., Age, Distance from Home) using StandardScaler.

Modeling

- **Logistic Regression**: Chosen for interpretability and suitability for binary classification.
- **Multicollinearity Check**: All VIF values < 2.0, confirming no significant collinearity.
- **Performance Metrics**: Accuracy, sensitivity, specificity, precision, and recall.

4. Visualizations

Training data between all the categorical columns and target variable to analyse how the categorical variables influence the target variable.

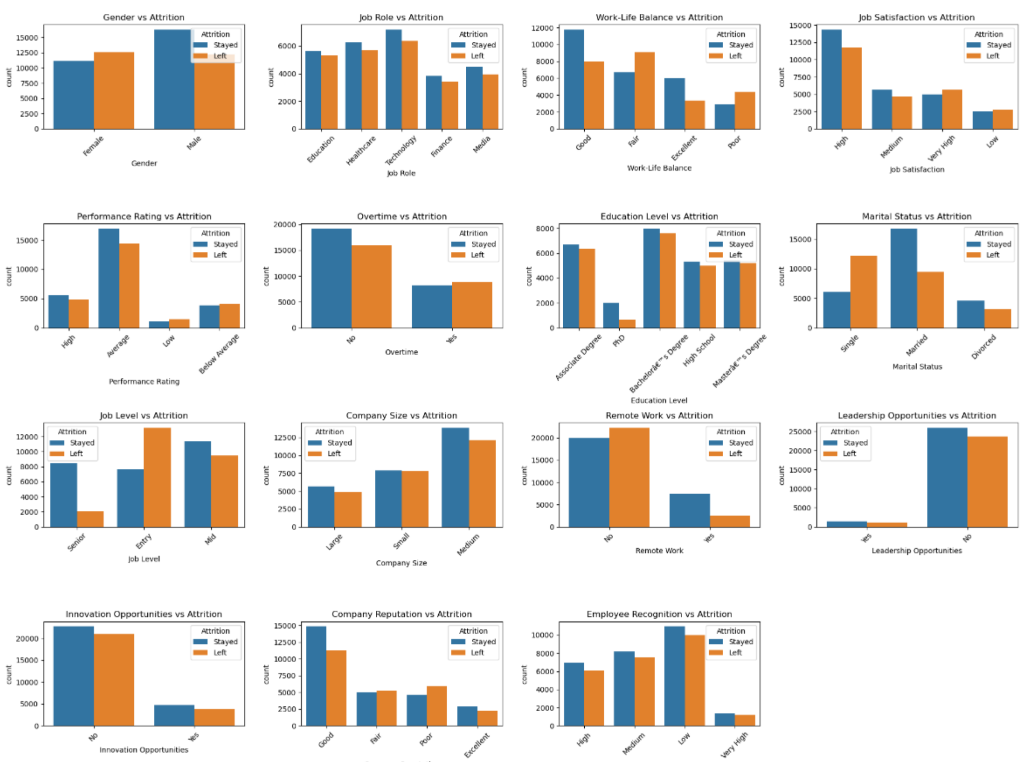
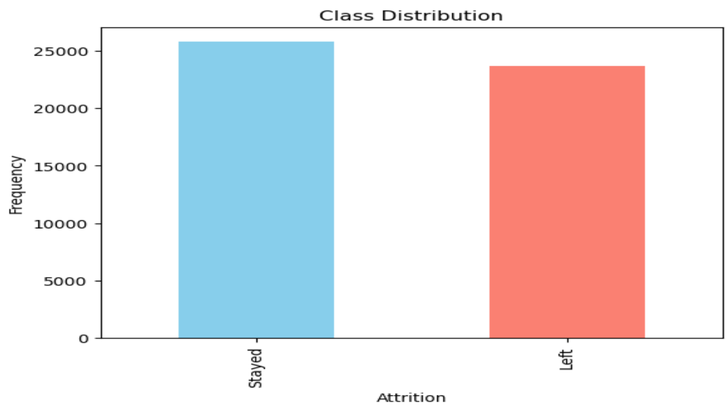
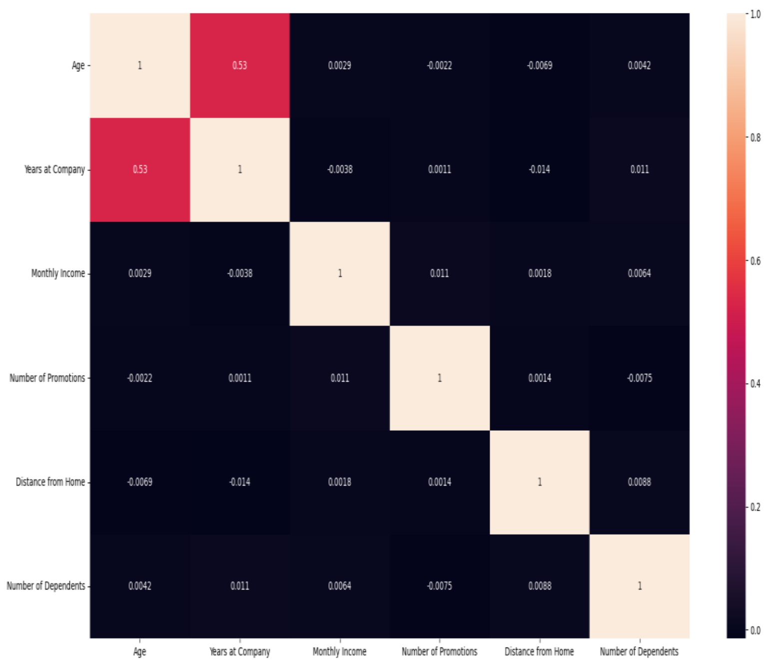


Figure 1: Class Distribution



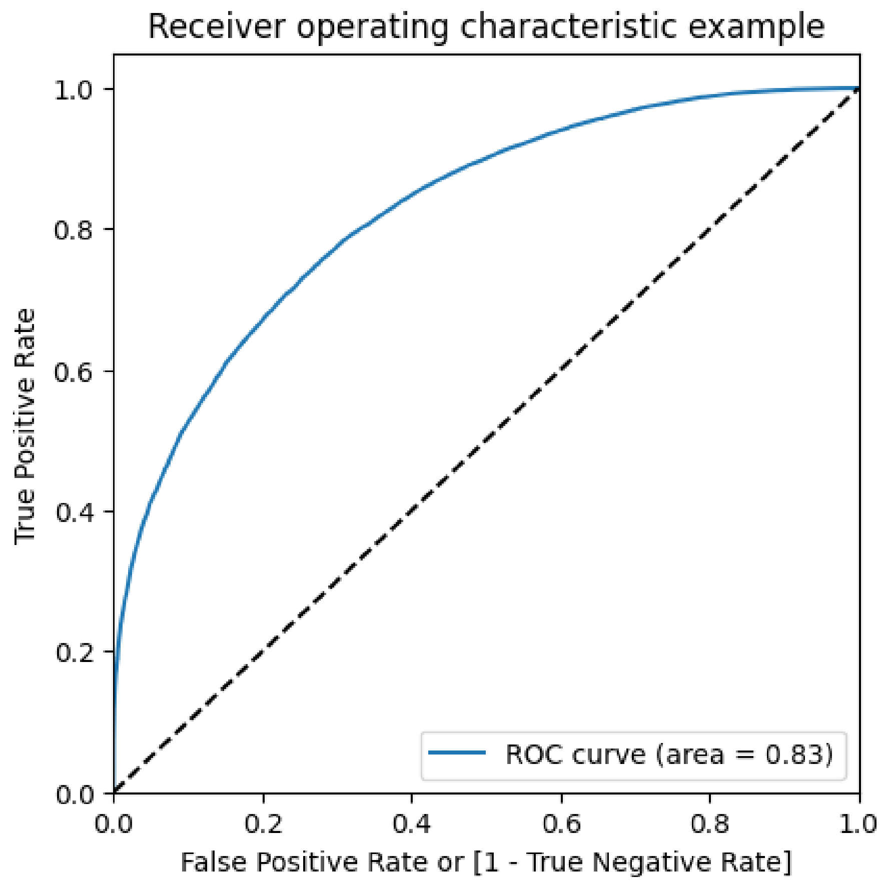
Insight: Severe class imbalance, with "Stayed" dominating the dataset.

Figure 2: Correlation Heatmap



- **Key Findings**:
  - Negative correlation between Work-Life Balance and attrition.
  - Monthly Income and Job Level positively correlated with retention.

Figure 4: ROC Curve



- **AUC**: 0.79, indicating moderate model discriminative power.

5. Key Insights

A. Retention Drivers

1. **Job Level**:
   - Senior employees are **12.6x more likely to stay** than junior staff.
2. **Remote Work**:
   - Remote workers show **5.6x higher retention odds**.
3. **Education**:
   - PhD holders are **4.4x more likely to stay**.

B. Attrition Drivers

1. **Work-Life Balance**:
   - Employees with "Poor" balance are **71% more likely to leave**.
2. **Marital Status**:
   - Single employees have **82% higher attrition risk**.

3. **Job Satisfaction**:
   - Paradoxically, **"Very High" satisfaction** correlates with attrition (38% higher risk).

C. Model Performance

Metric	Training	Validation
Accuracy	73.87%	73.57%
Sensitivity	75.33%	74.65%
Specificity	72.28%	72.40%
Precision	74.75%	74.59%

- **Consistency**: Minimal overfitting (training vs. test accuracy gap < 0.5%).

6. Actionable Outcomes

A. Retention Strategies

1. **Target High-Risk Groups**:
   - Implement mentorship programs for **single employees** and those in **low job levels**.
   - Offer flexible hours to improve **work-life balance** in high-attrition departments.
2. **Leverage Remote Work**:
   - Expand remote work policies to retain talent (e.g., hybrid models).
3. **Address the "Very High" Satisfaction Paradox**:
   - Conduct exit interviews to understand why highly satisfied employees leave.

B. Policy Recommendations

1. **Compensation Review**:
   - Benchmark salaries in roles with high attrition (e.g., Technology).
2. **Reputation Management**:
   - Improve employer branding through transparency and employee engagement initiatives.

C. Model Deployment

- **Predictive Monitoring**: Flag at-risk employees using the model for proactive HR interventions.
- **Threshold Adjustment**: Tune classification cutoff to prioritize sensitivity (reduce false negatives).

---

7. Conclusion

The logistic regression model achieved **73.5% accuracy** with balanced sensitivity (75%) and specificity (72%), identifying **job level, remote work, and marital status** as critical attrition drivers. While the model provides actionable insights, addressing the paradox of "Very High" job satisfaction and expanding remote work policies are key next steps.

Future Work:

- Explore ensemble models (e.g., Random Forest) for improved accuracy.
- Conduct qualitative research to validate counter-intuitive findings.

Appendix:

- **Confusion Matrix**:

	Predicted "Stayed"	**Predicted "Left"**
Actual "Stayed"	10,450	1,200
Actual "Left"	1,350	2,833

This report equips stakeholders with data-driven strategies to enhance retention and operational efficiency.