PREDICTING EMPLOYEE RETENTION REPORT

1.Problem Statement

1.1 Objective

This assignment aims to build a Logistic Regression model to predict binary outcomes, such as whether an employee will stay or leave. This task deepens understanding of logistic regression, its assumptions, implementation, and evaluation, especially for classification tasks.

1.2 Business Objective

A mid-sized technology company seeks to improve employee retention by identifying employees who are likely to stay. Traditionally, the focus was on managing turnover post-occurrence. This model aims to shift the focus toward proactive retention—understanding who is likely to remain and why. By analyzing demographic details, job satisfaction, performance, tenure, and more, this model provides HR with actionable insights to improve retention strategies and foster a stable, satisfied workforce.

1.3 Data Objective

The goal is to build a reliable model that predicts the likelihood of employee retention using the provided dataset. The challenges include:

- Identifying the most influential predictors of attrition.
- Handling missing or redundant features.
- Optimizing the model to guide HR in decision-making.
- Balancing sensitivity and specificity.

2.Dataset Overview

2.1 Dataset Information

- **Shape:** 74,610 rows X 24 columns
- **Key feature:** Age, Gender, Monthly income, Job role, Distance from home, Job satisfaction, Performance rating, Overtime, Job level, Education level, Attrition (target), etc.
- **Class distribution:**
  - 52.15% of employees stayed.
  - 47.85% of employees left.

2.2 Initial Data Exploration

- **Average age:** 38.53 years (± 12.08).
- **Monthly Income:** Mean = $7,344 (Range: $1,226 - $50,030).
- **Promotions:** Average = 0.83.
- **Missing values:** Found in distance from home (2.56%) and Company tenure (3.23%).

2.3 Data Types

- **Categorical:** Object (e.g., Gender, Job role, Work-life balance).
- **Numerical:** int64, float64 (e.g., Age, Income).

3.Methodology

3.1 Data Preprocessing

- **Missing values:** Rows with missing values were removed for data integrity.
- **Feature engineering:**
  - Redundant features (e.g., Employee ID, Employee recognition) were removed.

- ♠ Categorical variables were encoded using dummy variables.
- **Feature scaling:** Numerical features were standardized using a standard scaler.

3.2 Feature Selection

- **Statistical tests:**
  - Chi-square & Mann-Whitney U tests are confirmed feature significance.
  - Correlation analysis ruled out multicollinearity.
- **Recursive feature elimination (RFE) :**

  Selected 15 key predictors, including:

  - Gender- male, Work-life balance- fair/poor
  - Job satisfaction- low/very high, Performance rating- low/below average.
  - Overtime- yes, Education level- PhD, Remote work- yes.
  - Marital status- Single, Job level- mid/senior, Company reputation- poor/fair.

3.3 Model Building

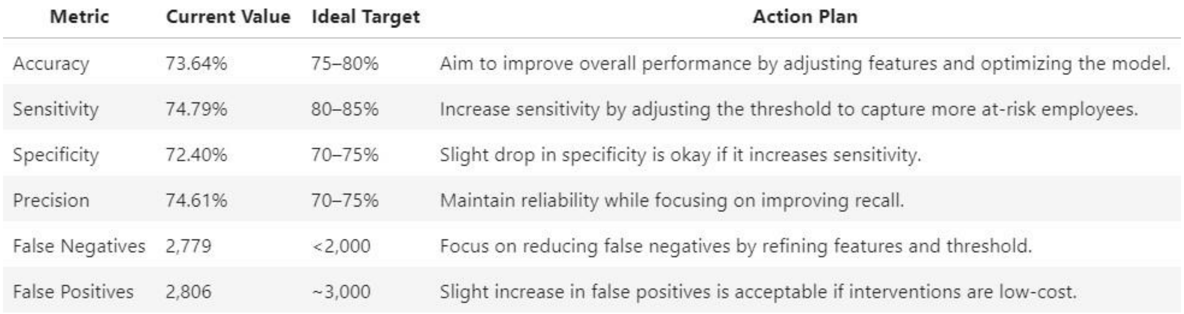
- **Logistic Regression:** The model was developed utilizing statistical models that included a constant term. Coefficients, p-values, and Variance Inflation Factors (VIFs) were assessed for their interpretation.

3.4 Threshold Optimization

- **ROC Curve AUC**: 0.8261 (indicating strong performance)
- **Youden's J Statistic**: Optimal threshold = 0.5513

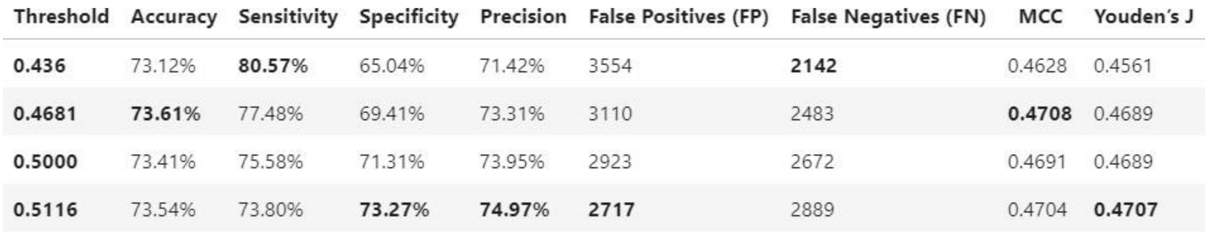
3.5 Model Evaluation

Iteration 1 (Threshold 0.5) -



Metric	Current Value	Ideal Target	Action Plan
Accuracy	73.64%	75–80%	Aim to improve overall performance by adjusting features and optimizing the model.
Sensitivity	74.79%	80–85%	Increase sensitivity by adjusting the threshold to capture more at-risk employees.
Specificity	72.40%	70–75%	Slight drop in specificity is okay if it increases sensitivity.
Precision	74.61%	70–75%	Maintain reliability while focusing on improving recall.
False Negatives	2,779	<2,000	Focus on reducing false negatives by refining features and threshold.
False Positives	2,806	~3,000	Slight increase in false positives is acceptable if interventions are low-cost.

Experimenting with different thresholds:



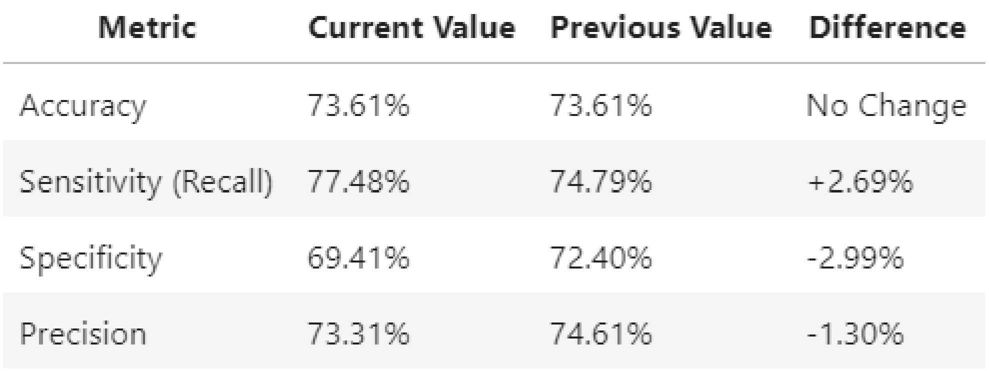
Threshold	Accuracy	Sensitivity	Specificity	Precision	False Positives (FP)	False Negatives (FN)	MCC	Youden's J
0.436	73.12%	**80.57%**	65.04%	71.42%	3554	**2142**	0.4628	0.4561
0.4681	**73.61%**	77.48%	69.41%	73.31%	3110	2483	**0.4708**	0.4689
0.5000	73.41%	75.58%	71.31%	73.95%	2923	2672	0.4691	0.4689
0.5116	73.54%	73.80%	**73.27%**	**74.97%**	**2717**	2889	0.4704	**0.4707**

1. **Threshold = 0.4681**:
    - This is the **best balance**: it catches **77.48%** of employees who are likely to leave (Sensitivity) while avoiding too many false alarms (**69.41%** Specificity).
    - **MCC (0.4708)** shows the model is performing well overall.
2. **Threshold = 0.5000** (default):
    - Slightly lower Sensitivity (**75.58%**) means it might miss some employees who would leave.
    - Slightly better Specificity (**71.31%**) reduces false alarms but might miss some at-risk employees.
3. **Threshold = 0.5116**:
    - Best **Precision** (74.97%) and **Specificity** (73.27%) mean fewer false alerts, but it misses more at-risk employees (**73.80% Sensitivity**).
4. **Threshold = 0.436**:
    - Highest **Sensitivity** (80.57%) catches most leavers, but it causes too many false alarms (**65.04% Specificity**), leading to wasted efforts.

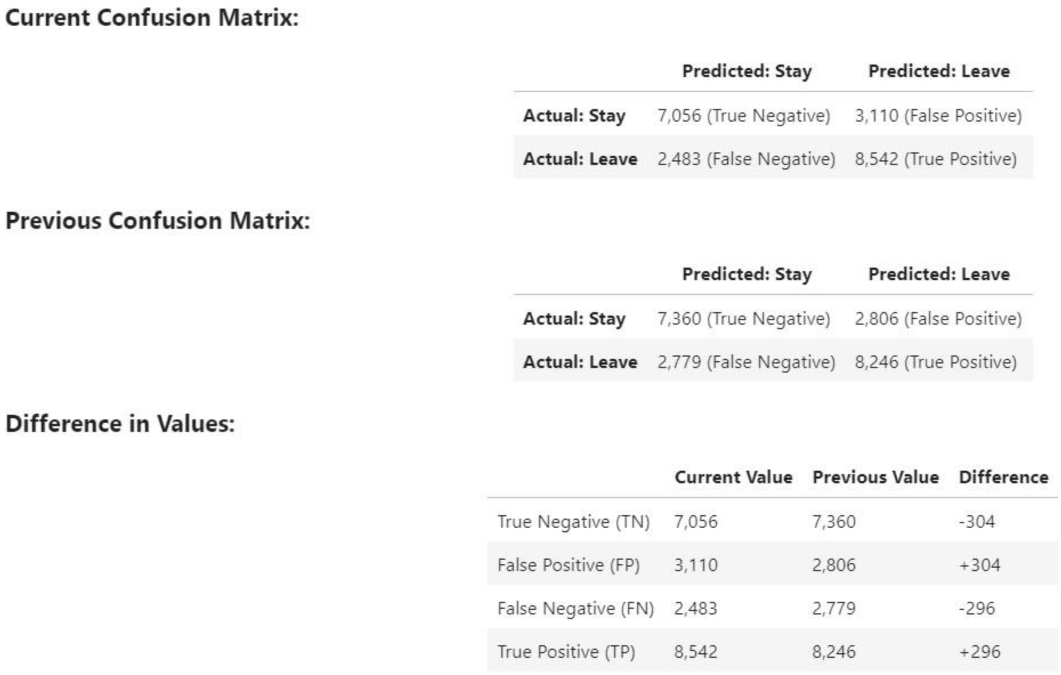
Using a **threshold of 0.4681** because it strikes the best balance between:

- **Capturing at-risk employees**: It finds **77.48%** of those who will leave.
- **Reducing false positives**: False alarms are kept at manageable levels.
- **Strong overall performance**: With an **MCC of 0.4708**, the model stays balanced and stable.

Iteration 2



Metric	Current Value	Previous Value	Difference
Accuracy	73.61%	73.61%	No Change
Sensitivity (Recall)	77.48%	74.79%	+2.69%
Specificity	69.41%	72.40%	-2.99%
Precision	73.31%	74.61%	-1.30%



Current Confusion Matrix:

	Predicted: Stay	Predicted: Leave
Actual: Stay	7,056 (True Negative)	3,110 (False Positive)
Actual: Leave	2,483 (False Negative)	8,542 (True Positive)

Previous Confusion Matrix:

	Predicted: Stay	Predicted: Leave
Actual: Stay	7,360 (True Negative)	2,806 (False Positive)
Actual: Leave	2,779 (False Negative)	8,246 (True Positive)

Difference in Values:

	Current Value	Previous Value	Difference
True Negative (TN)	7,056	7,360	-304
False Positive (FP)	3,110	2,806	+304
False Negative (FN)	2,483	2,779	-296
True Positive (TP)	8,542	8,246	+296

Key Improvements:

- 296 fewer False Negatives.
- 296 more True Positives.
- Trade-off: 304 more False Positives (acceptable for HR interventions).

4.Key Insights

4.1 Factors Contributing to Attrition

- **High Risk**:
  - Low job satisfaction.
  - Below-Average performance.
  - Overtime work.
  - Low education.
  - Single marital status.
  - Entry-level roles.
  - Remote work.
  - Poor company reputation.

4.2 Feature Insights

- **Positive Coefficients** (increase attrition probability):
  - ♠ Gender- Male, Remote Work- Yes, Senior Job Level, PhD
- **Negative Coefficients** (decrease attrition probability):
  - ♠ Poor Work-Life Balance, Low Job Satisfaction, Single Status, Poor Reputation.

4.3 Insights

- **Age**: The mean age of employees is approximately 38.5 years, with a standard deviation of 12.08, indicating a relatively diverse age range. The distribution appears roughly normal with a slight platykurtic shape.

- **Attrition**: 47.85% of employees left the company, while 52.15% remained, indicating a relatively balanced distribution of attrition.

- **Company Reputation**: 50% of respondents rated the company as “Good,” while 20.25% rated it as “Poor.” The chi-square test results indicate that the company’s reputation significantly impacts customer retention.

- **Company Size**: 50.1% of employees work in medium-sized companies, and 29.9% in small companies. The chi-square test suggests that company size is related to attrition, with smaller companies having higher turnover.

- **Company Tenure**: Average tenure is 56 months, with a standard deviation of 25.4 months. The tenure distribution exhibits a slight positive skew, with most employees having medium tenure.

- **Distance from Home**: The average distance from home is 49.89 miles, showing a normal distribution.

- **Education Level**: The majority of employees have a Bachelor's Degree (30%), followed by an Associate Degree (24.86%). Education level is significantly linked to attrition.

- **Gender**: 55% of employees are male, with 45% female. Gender significantly affects attrition.

- **Innovation Opportunities**: 83.86% have no innovation opportunities, and this is crucially related to attrition.

- **Job Level**: Most employees are in Mid or Entry-level roles, with job level strongly linked to attrition.

- **Job Satisfaction**: 50.18% of employees have high job satisfaction, and this factor is closely corelated to attrition.

- **Leadership Opportunities**: A majority (95.11%) do not have leadership opportunities, and this is significantly related to attrition.

- **Marital Status**: Most employees are married (50.3%), with marital status significantly influencing attrition.

- **Monthly Income**: The average monthly income is $7,347.36, with a high skewness, indicating some high-income outliers.

- **Number of Dependents**: The average number of dependents is 1.66, with a slight positive skew.

- **Number of Promotions**: The average number of promotions is 0.83, with a right-skewed distribution.

- **Overtime**: 67.54% do not work overtime, with overtime status significantly linked to attrition.

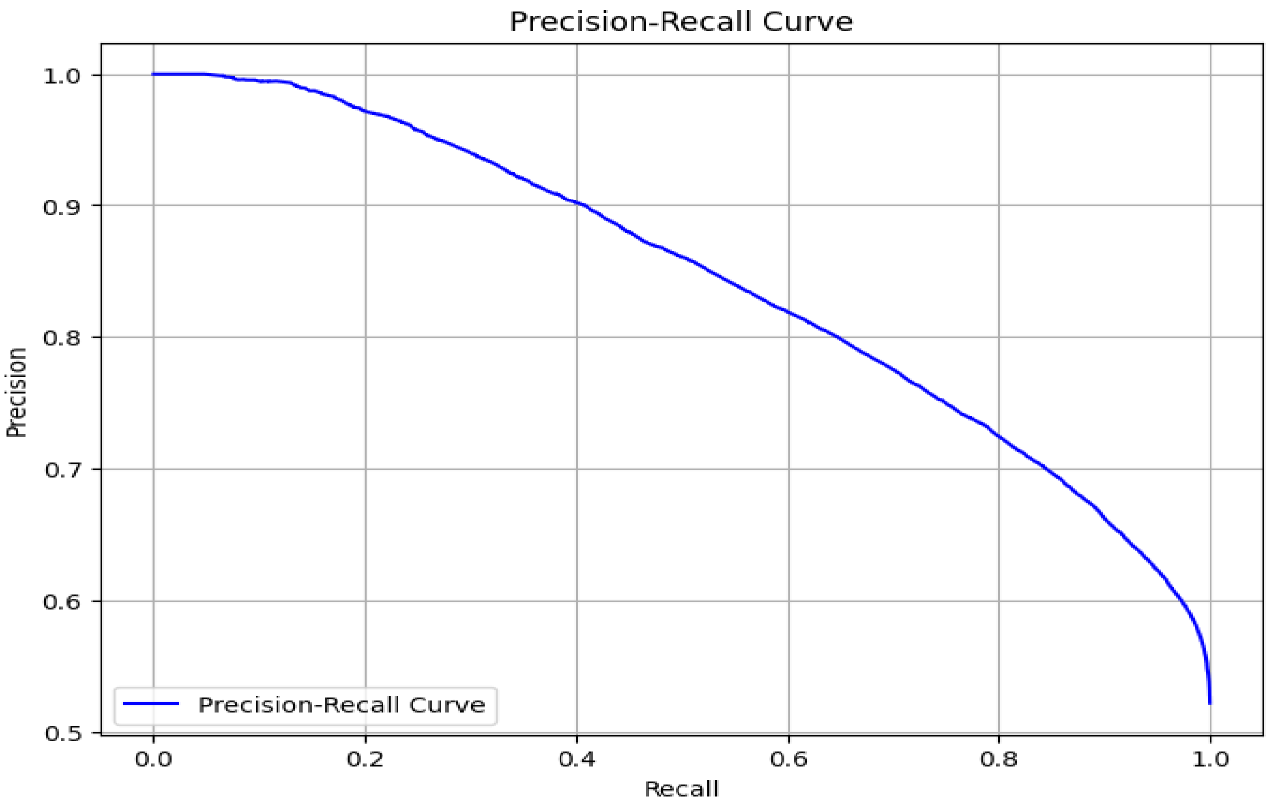
- **Performance Rating**: Most employees have an average performance rating (59.99%), with lower performance strongly linked to attrition.

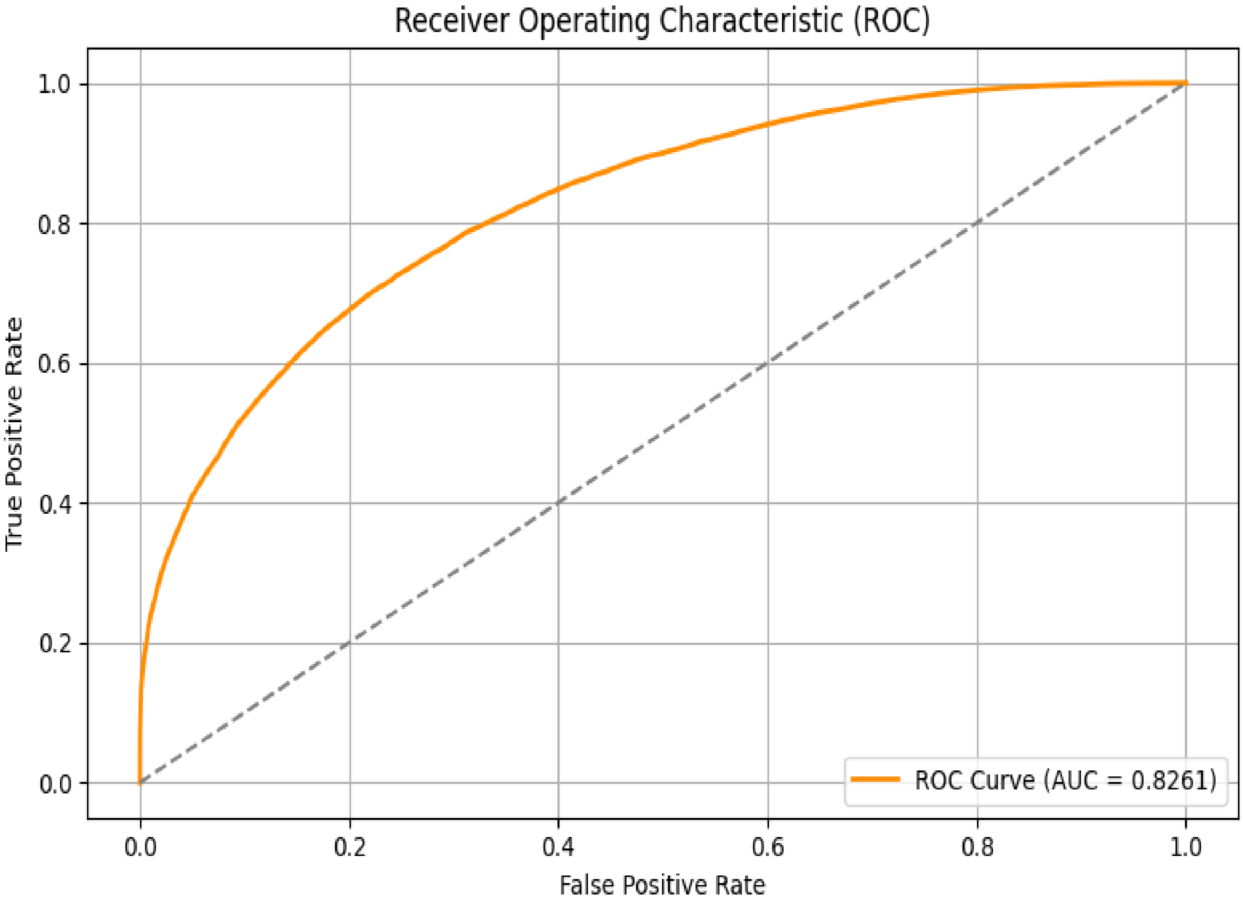
- **Remote Work**: 82.24% do not work remotely, with remote work strongly linked to attrition.

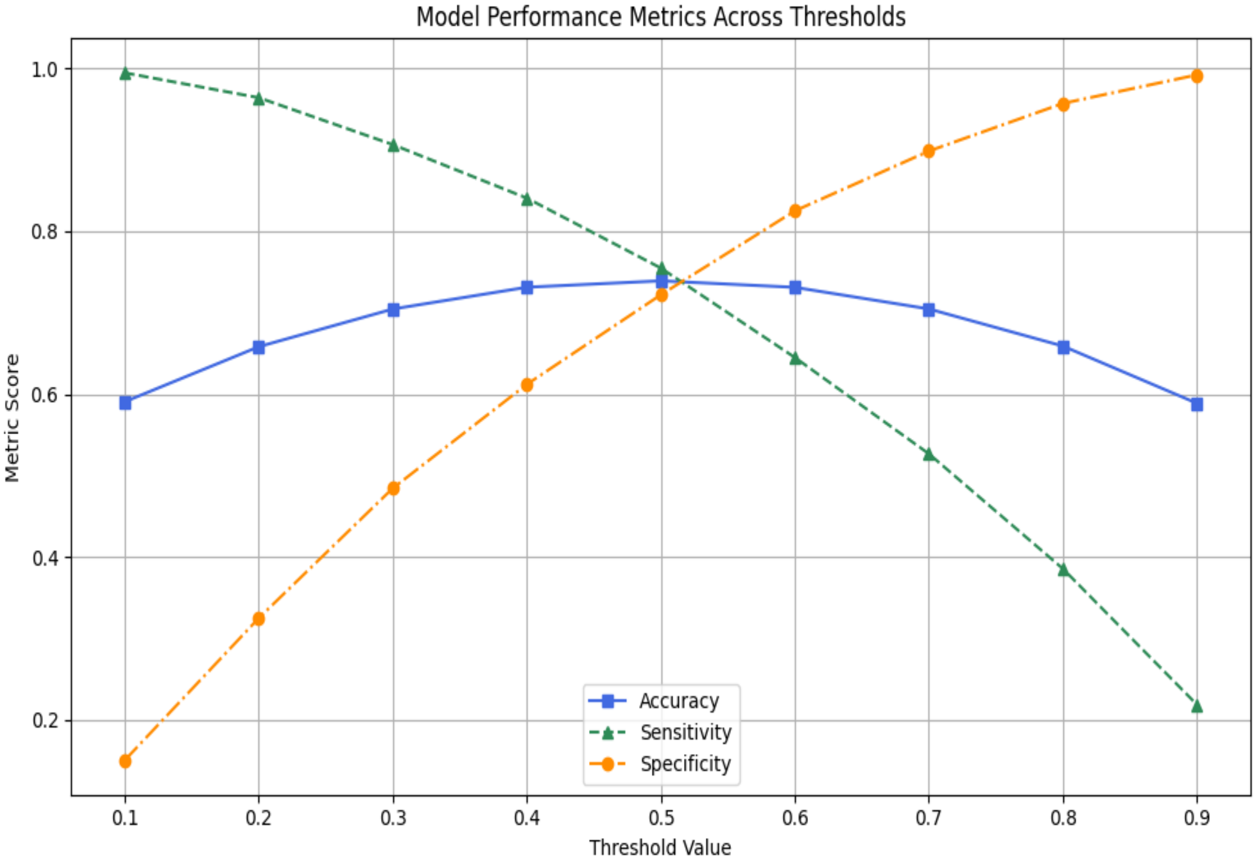
5.Techniques Used

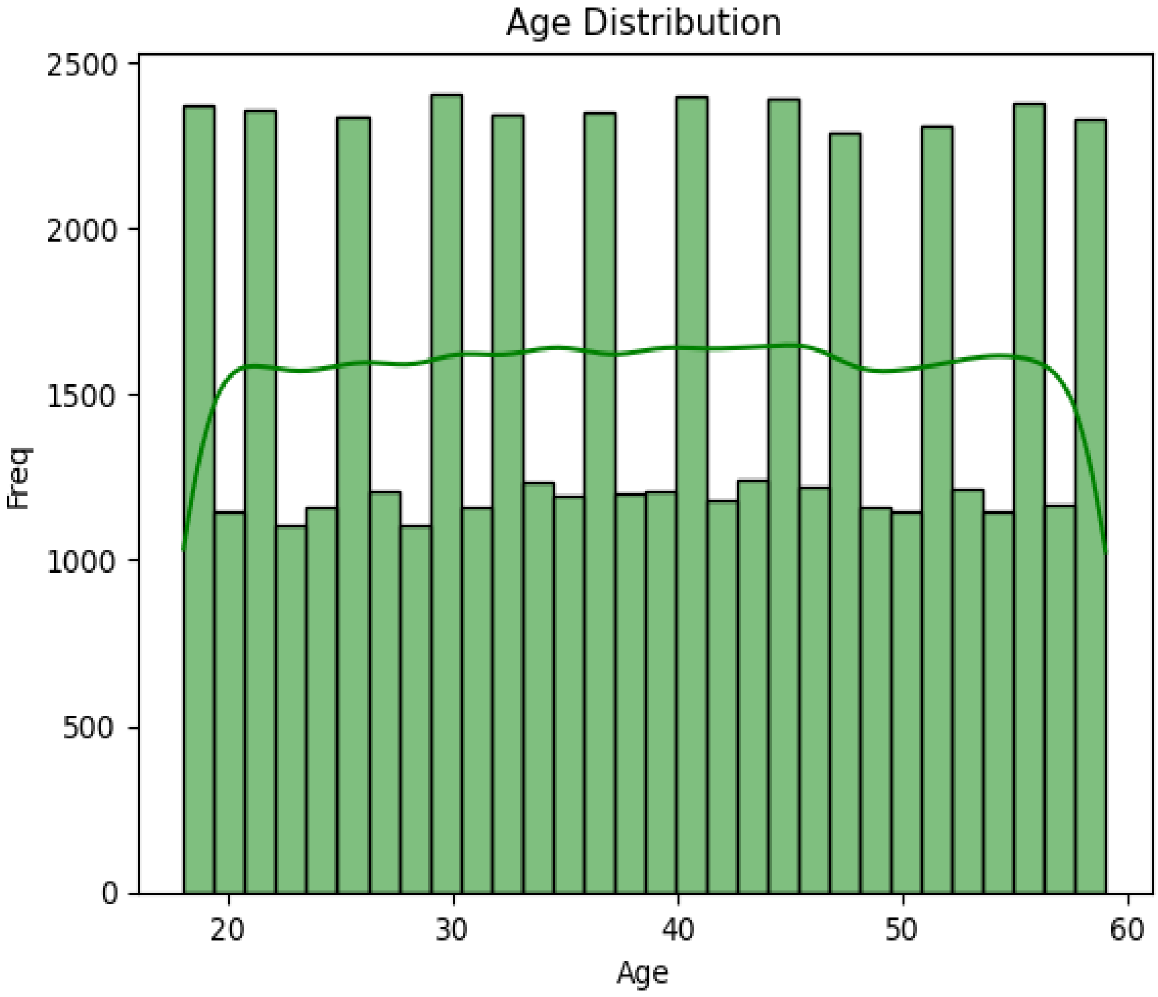
- **EDA**: Summary statistics, missing value handling, correlation matrix.
- **Feature Engineering**: Dummy variables, feature scaling, feature elimination.
- **Feature Selection**: Variance threshold, mutual information, RFE.
- **Modeling**: Logistic Regression.
- **Threshold Optimization**: ROC analysis, Youden’s J.
- **Evaluation Metrics**: Accuracy, Sensitivity, Specificity, Precision, Confusion Matrix.

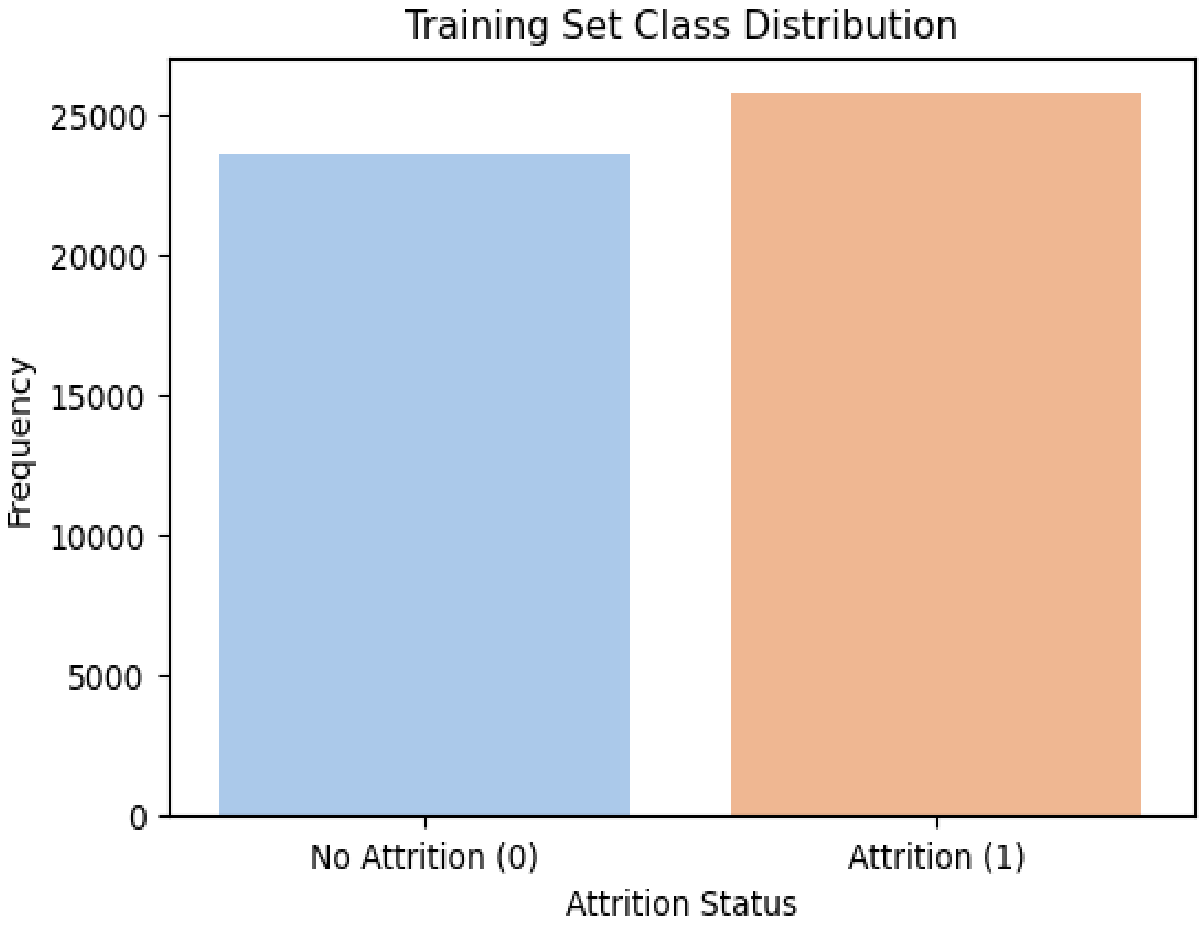
6. Visualizations and Plots

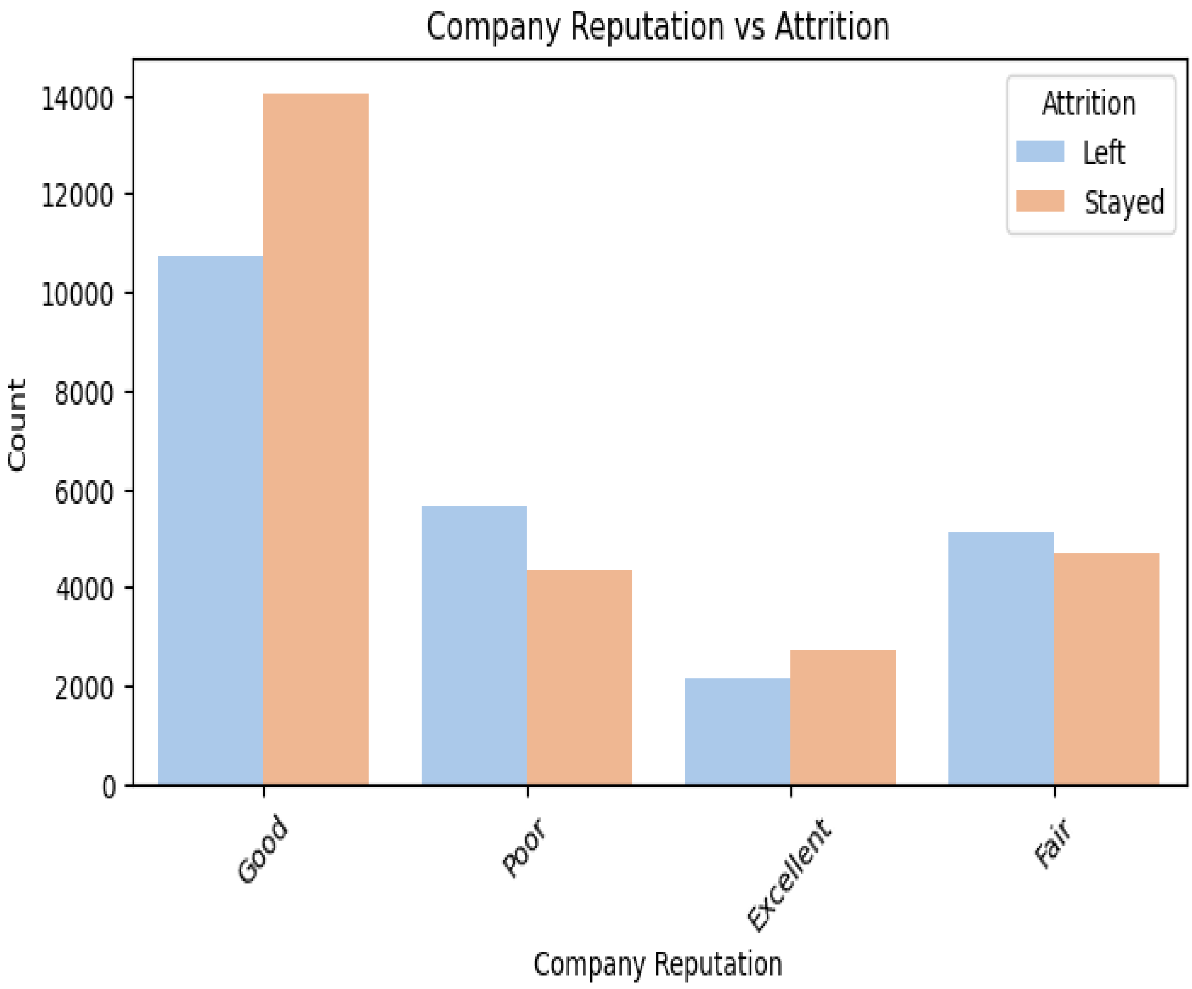


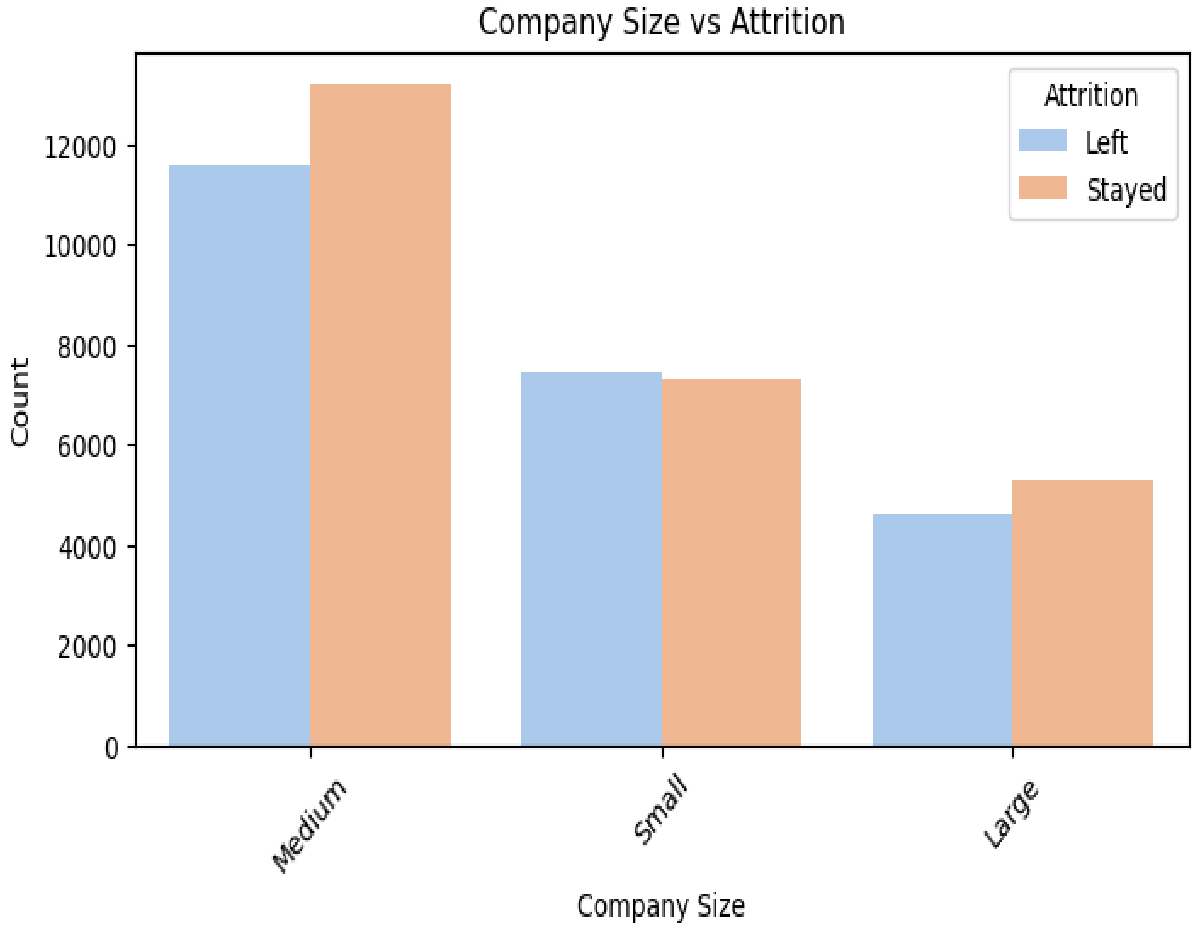


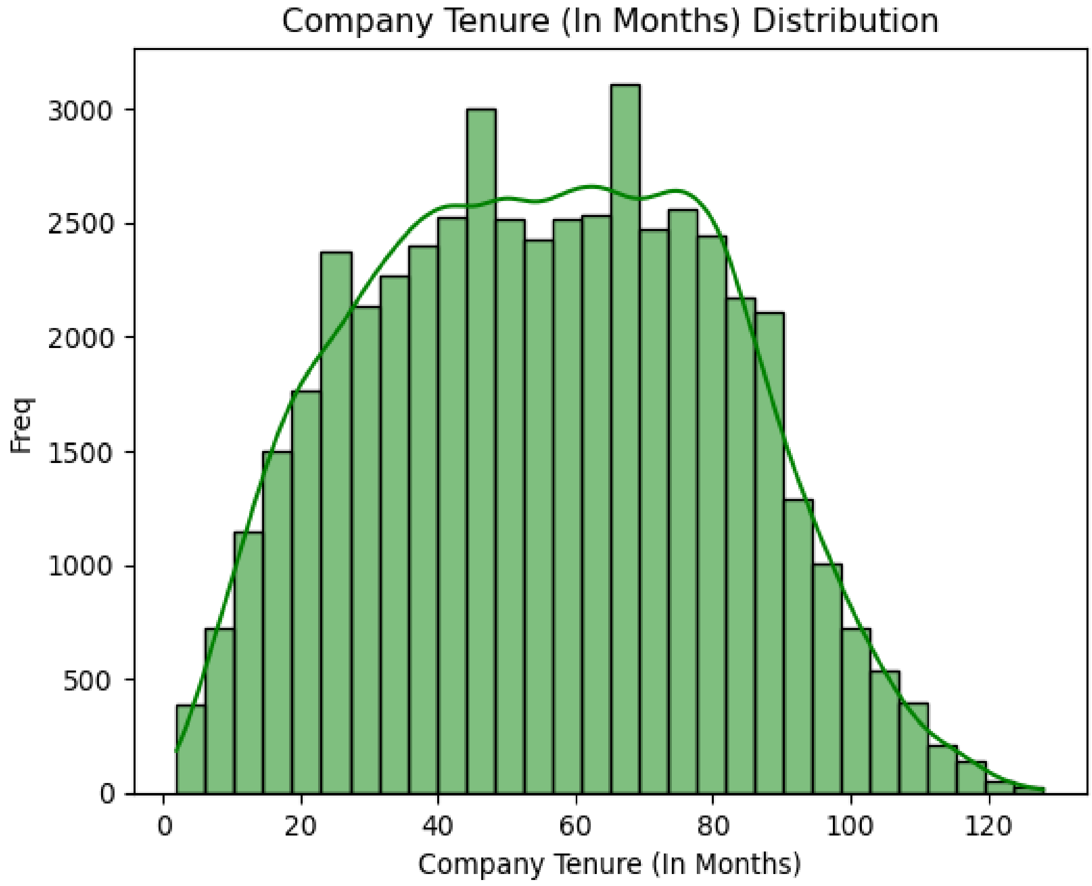

Model Performance Metrics Across Thresholds
Metric Score
1.0
0.8
0.6
0.4
0.2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Threshold Value
Accuracy
Sensitivity
Specificity

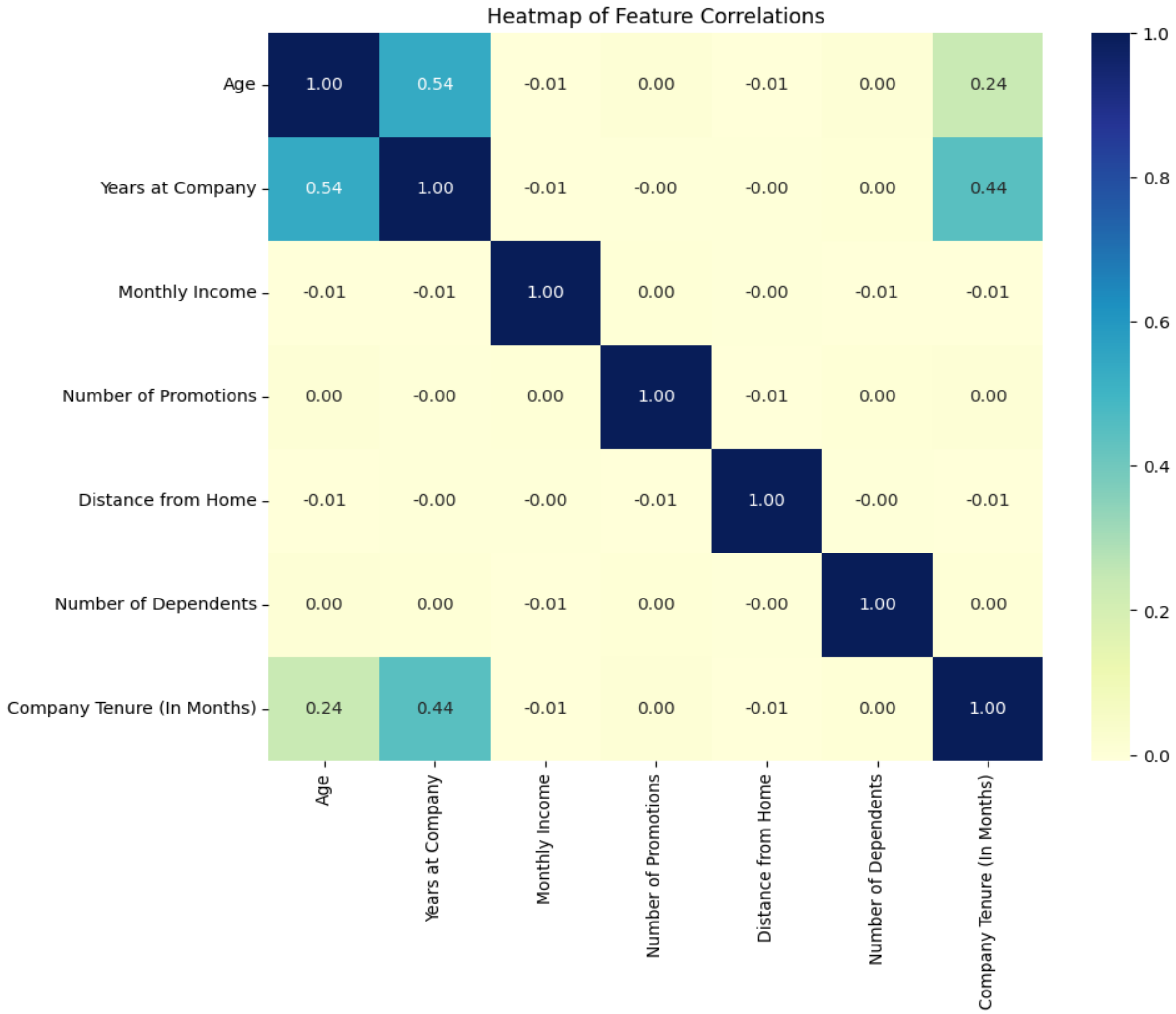

Age Distribution
Freq
2500
2000
1500
1000
500
0
20
30
40
50
60
Age

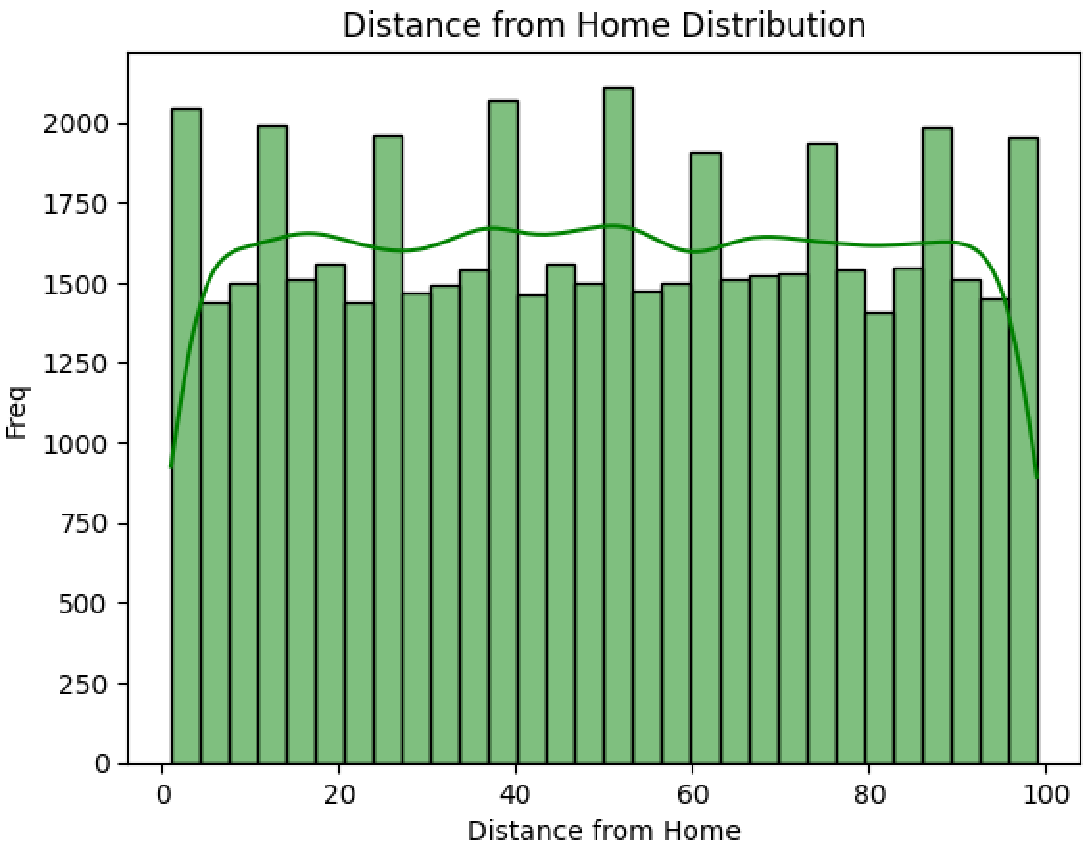

Training Set Class Distribution
Frequency
25000
20000
15000
10000
5000
0
No Attrition (0)
Attrition (1)
Attrition Status

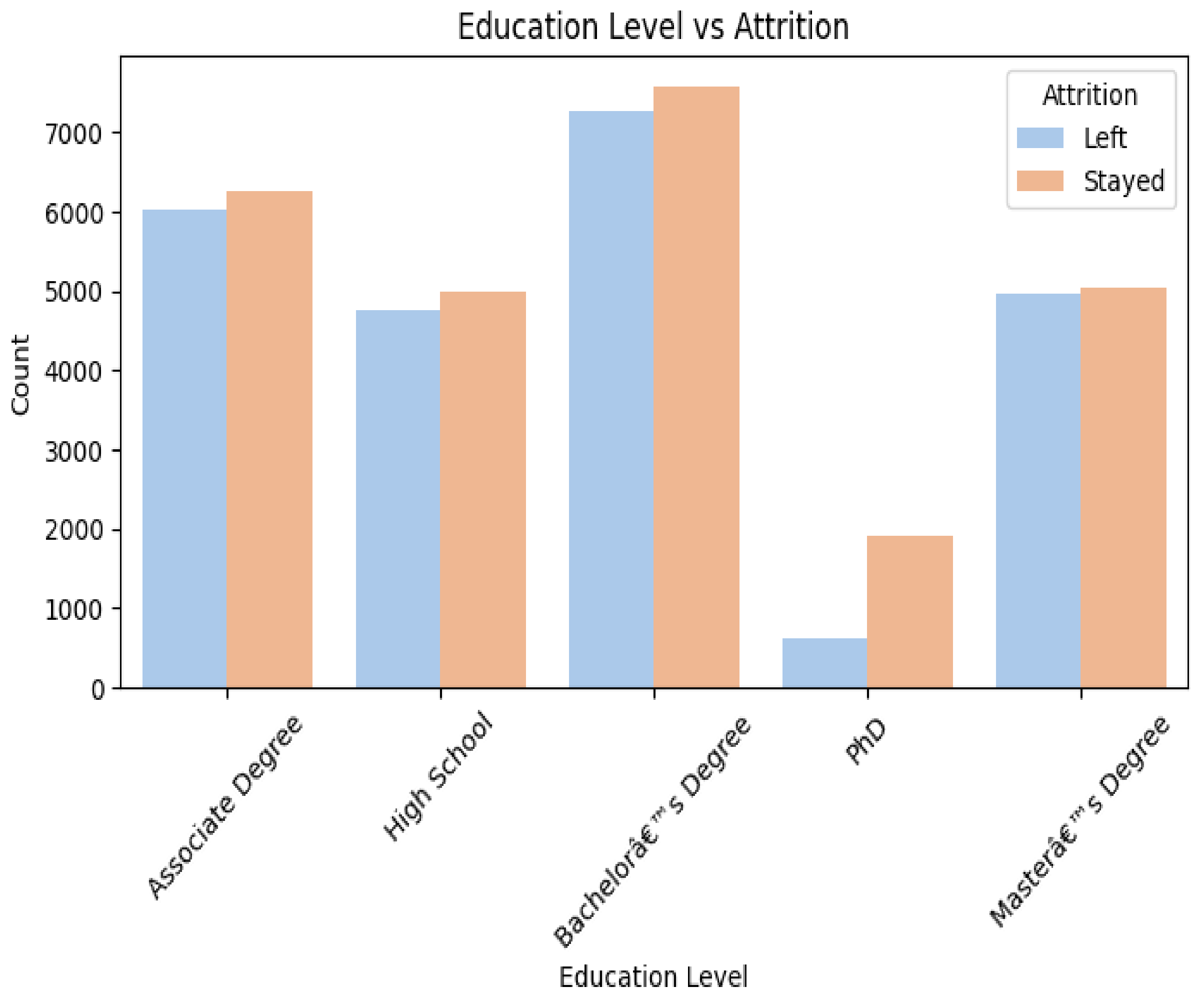

Company Reputation vs Attrition
Count
14000
12000
10000
8000
6000
4000
2000
0
Attrition
Left
Stayed
Good
Poor
Excellent
Fair
Company Reputation

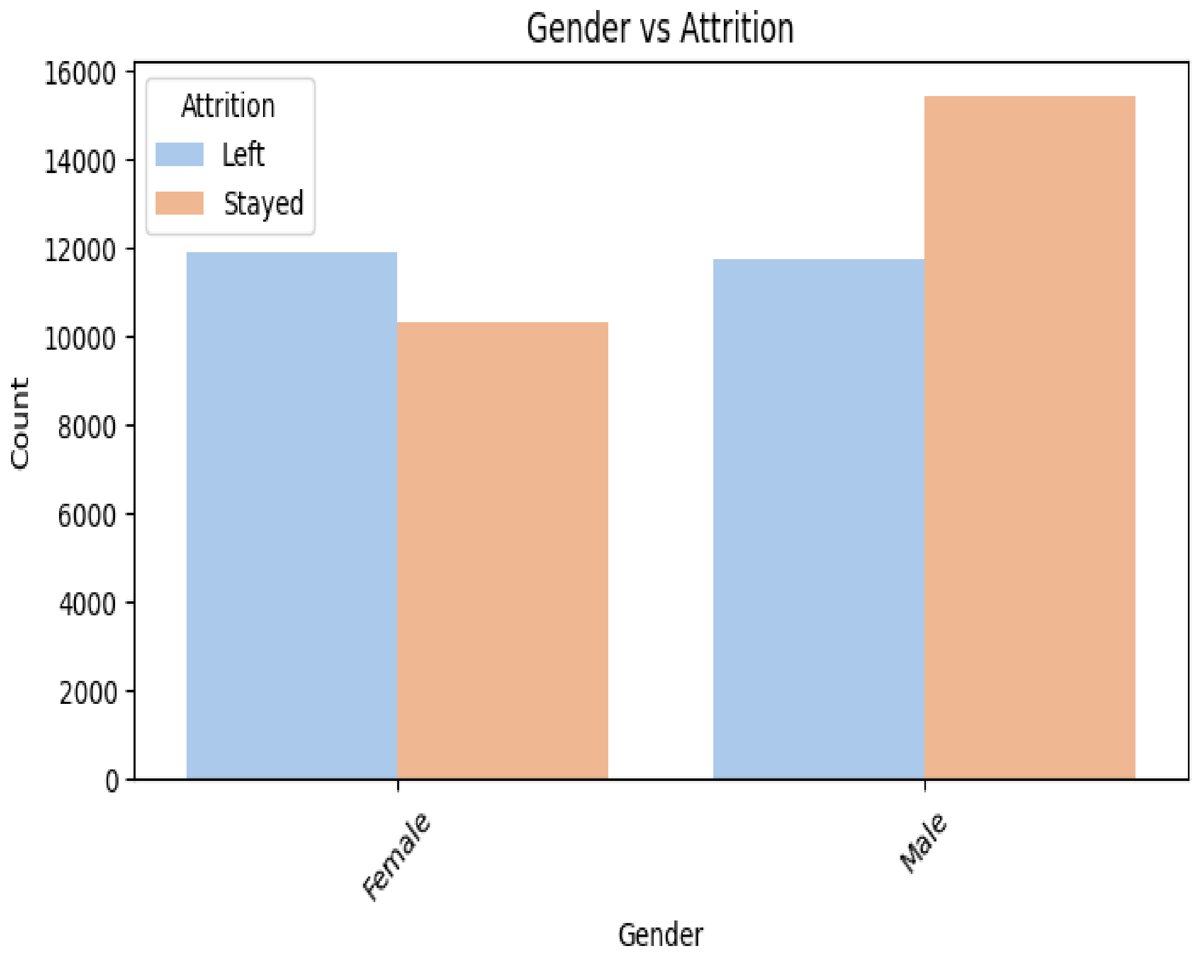

Company Size vs Attrition
Count
12000
10000
8000
6000
4000
2000
0
Medium
Small
Large
Company Size
Attrition
Left
Stayed

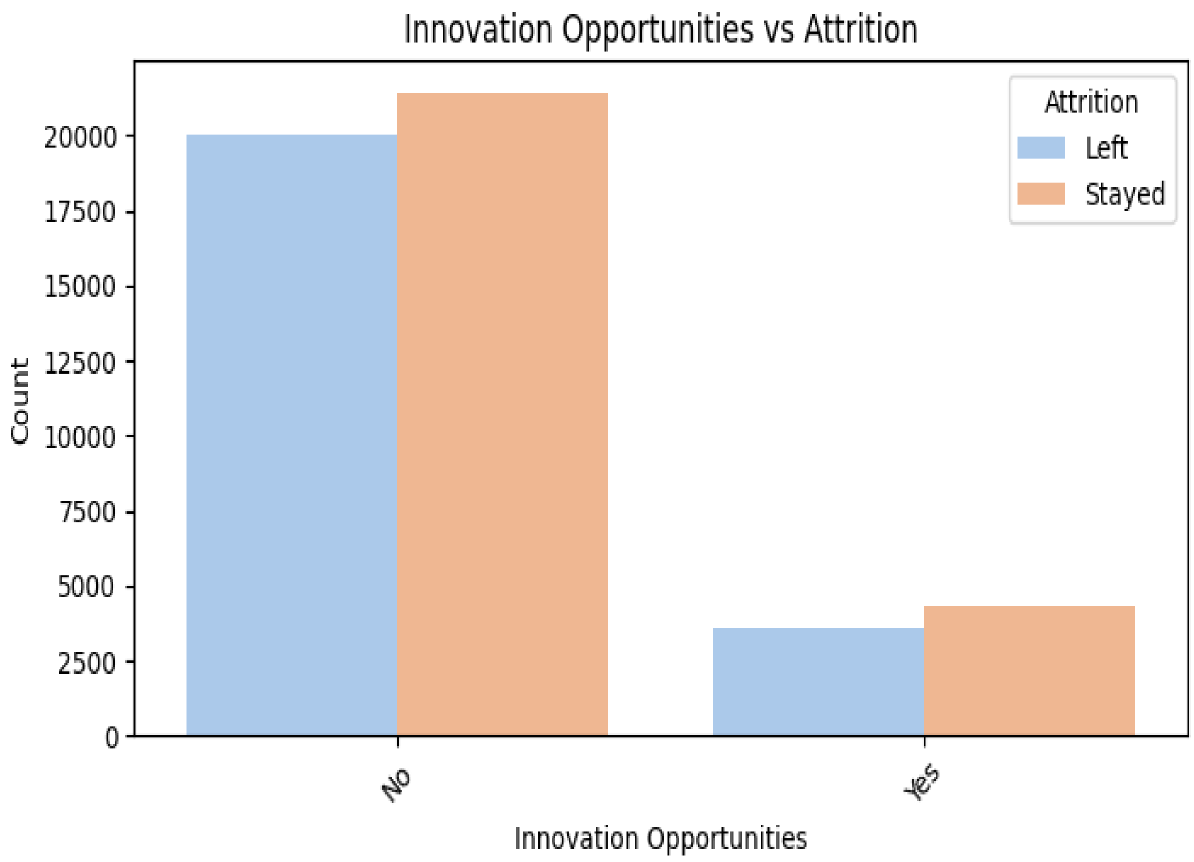

Company Tenure (In Months) Distribution
Freq
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100
120
Company Tenure (In Months)

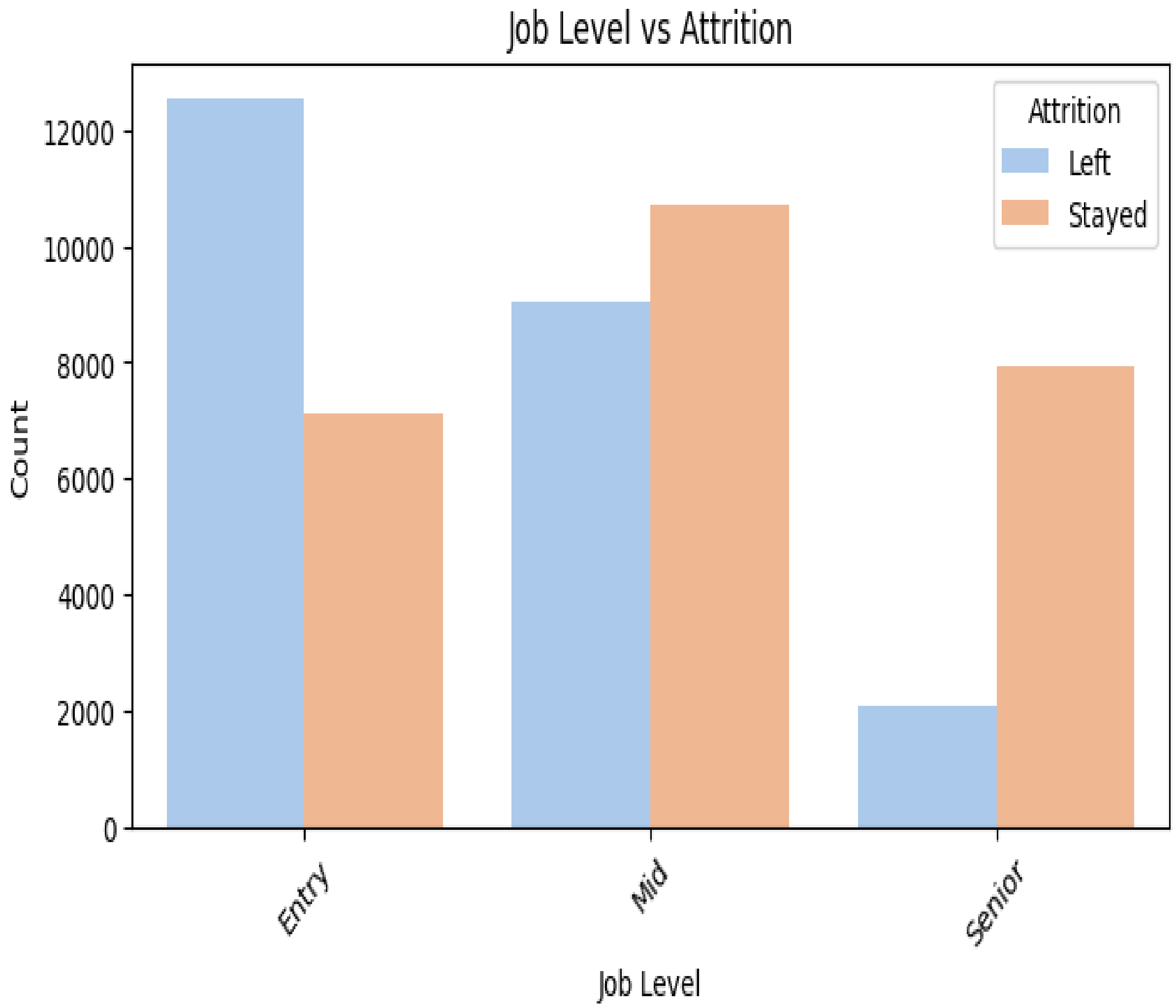

Heatmap of Feature Correlations
Age
Years at Company
Monthly Income
Number of Promotions
Distance from Home
Number of Dependents
Company Tenure (In Months)
1.00 0.54 -0.01 0.00 -0.01 0.00 0.24
0.54 1.00 -0.01 -0.00 -0.00 0.00 0.44
-0.01 -0.01 1.00 0.00 -0.00 -0.01 -0.01
0.00 -0.00 0.00 1.00 -0.01 0.00 0.00
-0.01 -0.00 -0.00 -0.01 1.00 -0.00 -0.01
0.00 0.00 -0.01 0.00 -0.00 1.00 0.00
0.24 0.44 -0.01 0.00 -0.01 0.00 1.00
1.0
0.8
0.6
0.4
0.2
0.0

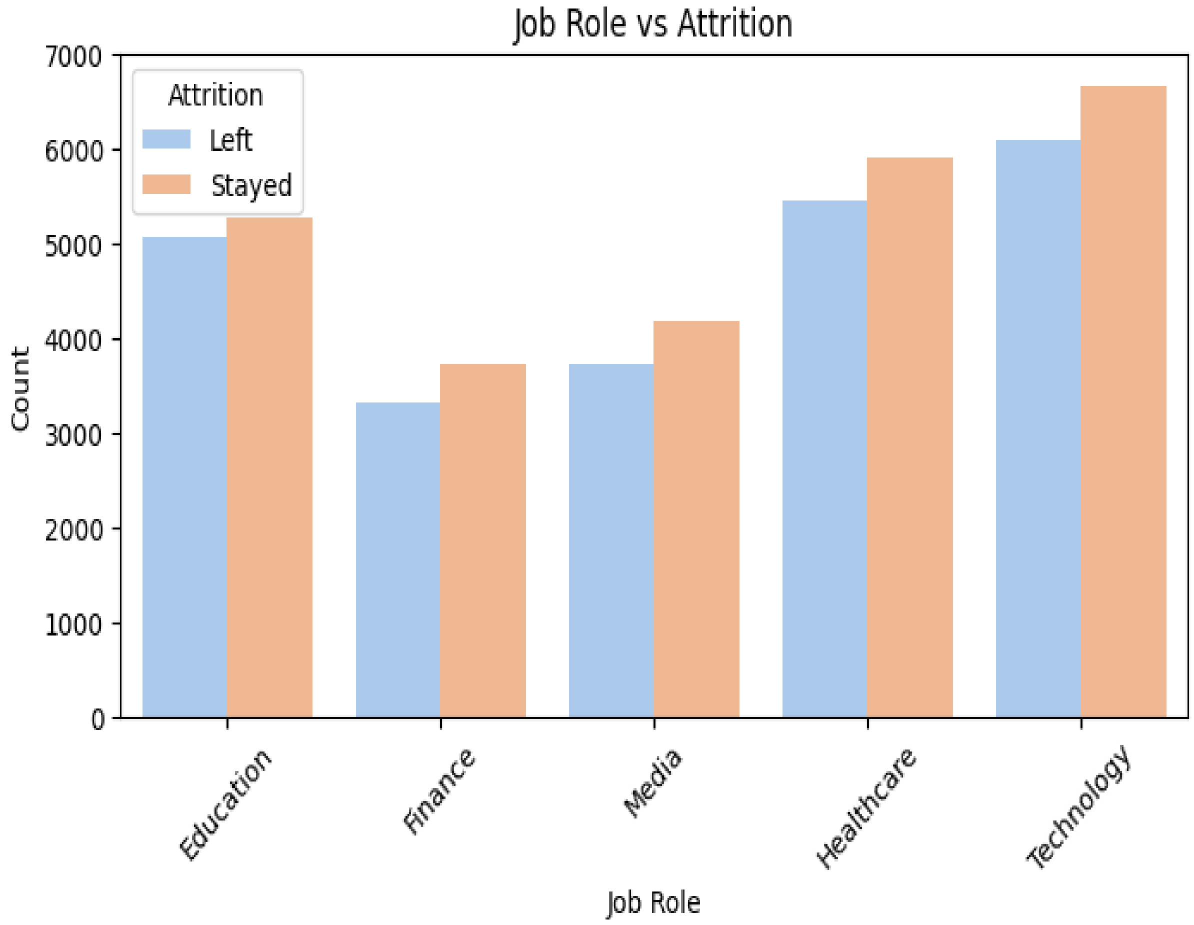

Distance from Home Distribution
Freq
0
250
500
750
1000
1250
1500
1750
2000
0
20
40
60
80
100
Distance from Home

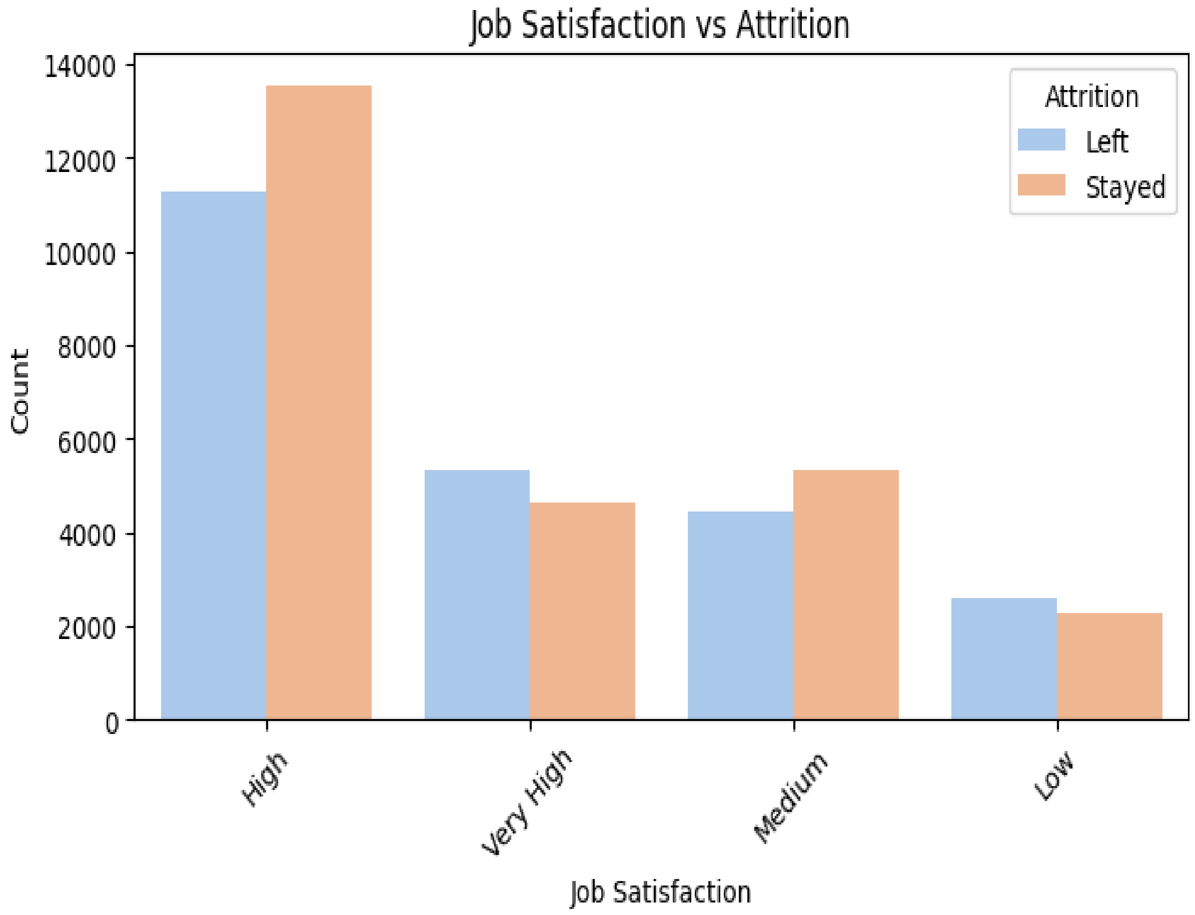

Education Level vs Attrition
Attrition
Left
Stayed
Count
7000
6000
5000
4000
3000
2000
1000
0
Associate Degree
High School
Bachelorâ€™s Degree
PhD
Masterâ€™s Degree
Education Level

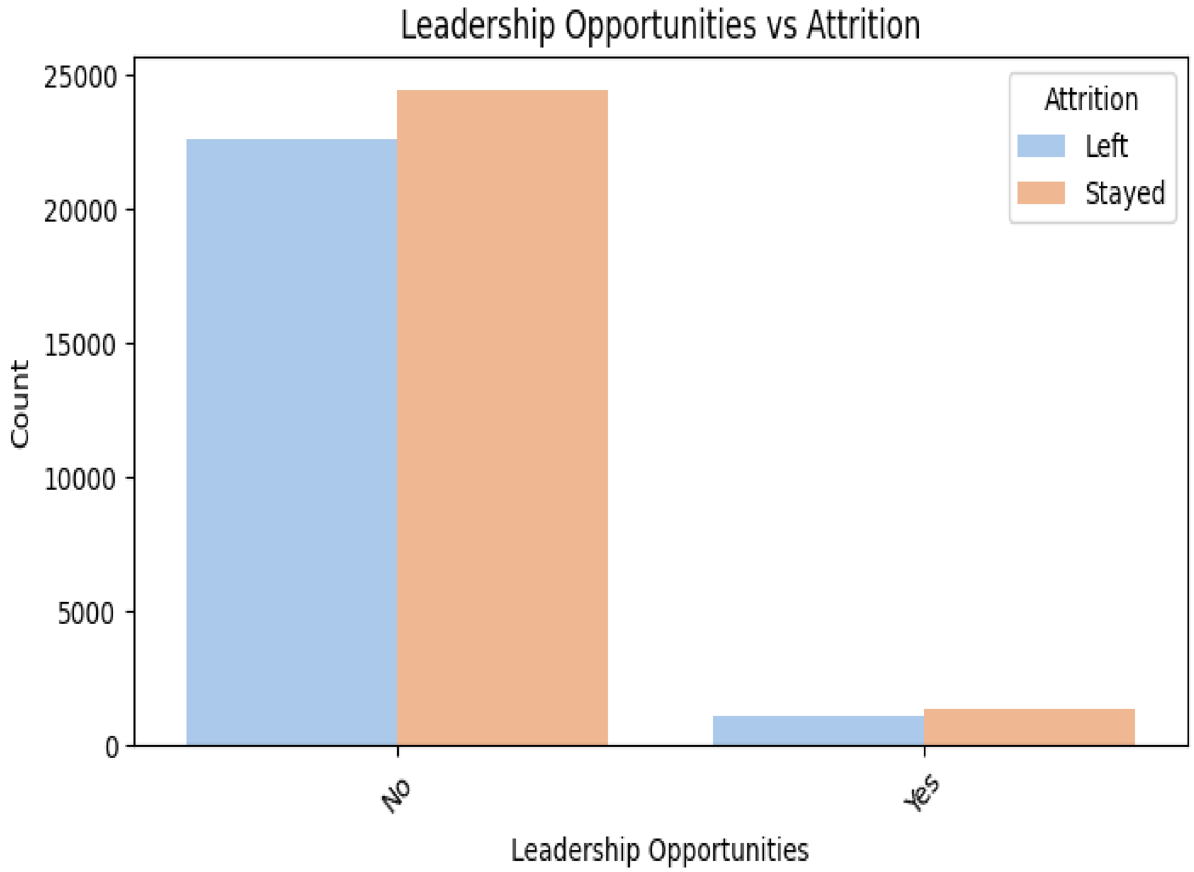

Gender vs Attrition
Attrition
Left
Stayed
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
Female
Male
Gender

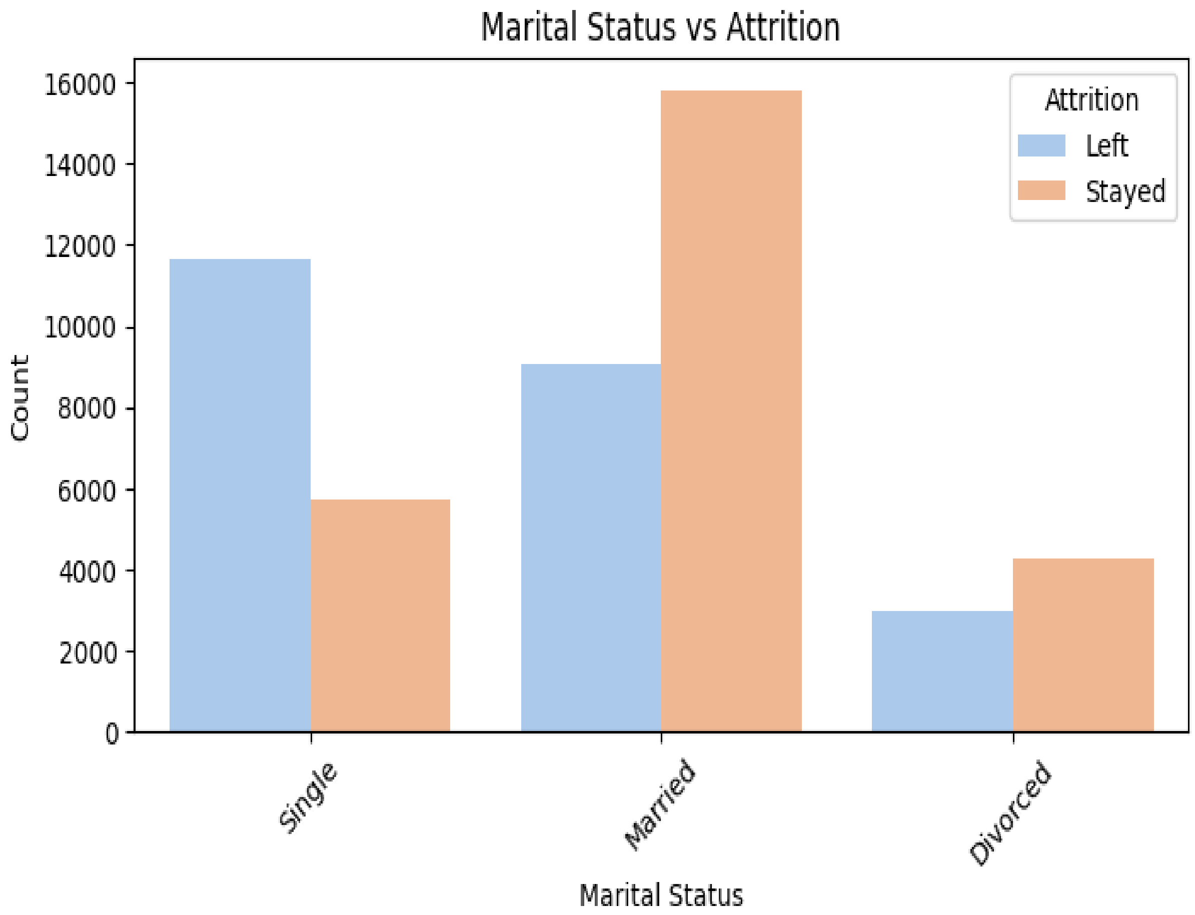

Innovation Opportunities vs Attrition
Count
0
2500
5000
7500
10000
12500
15000
17500
20000
No
Yes
Innovation Opportunities
Attrition
Left
Stayed

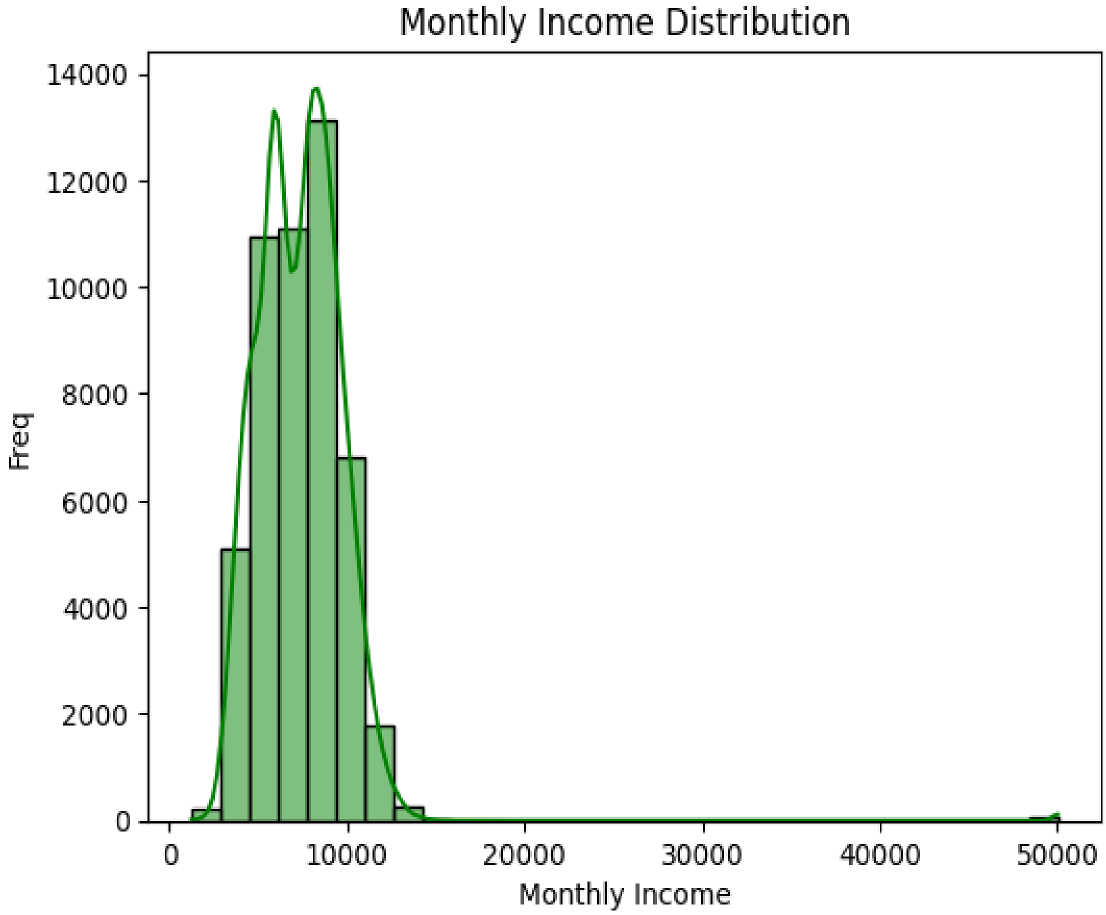

Job Level vs Attrition
Count
0
2000
4000
6000
8000
10000
12000
Entry
Mid
Senior
Job Level
Attrition
Left
Stayed

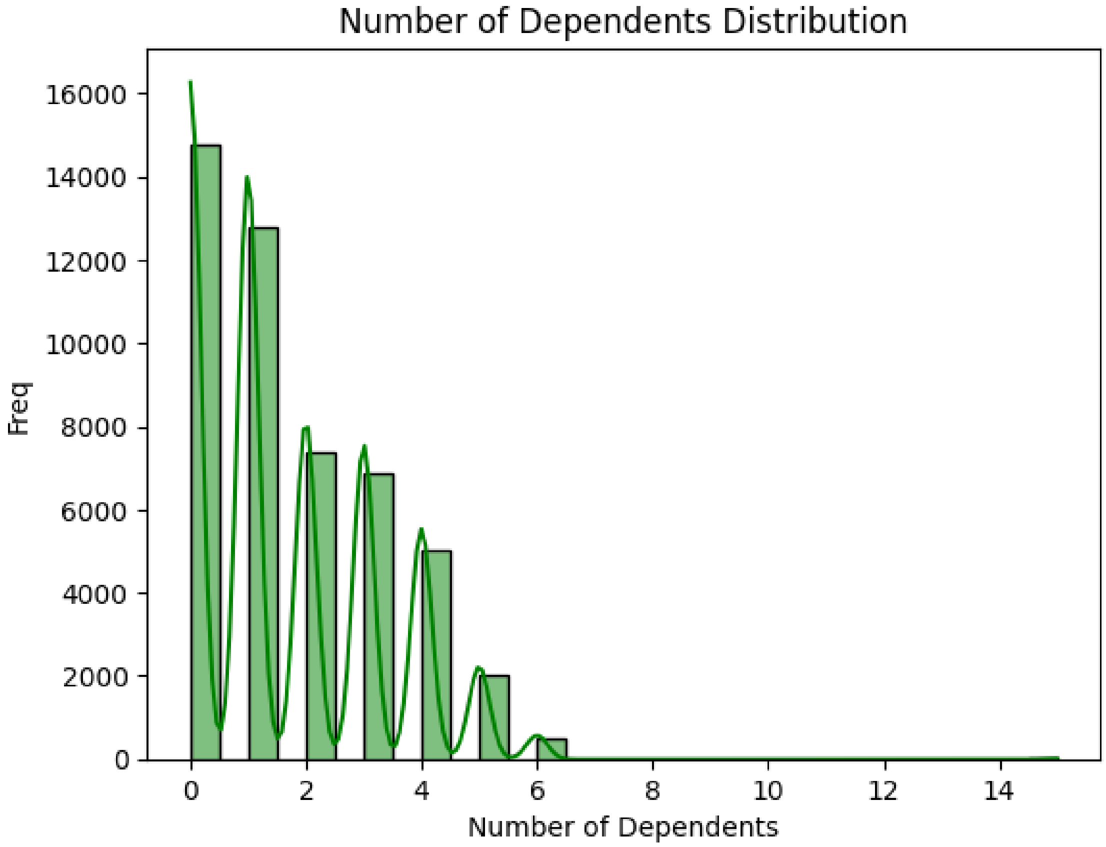

Job Role vs Attrition
Attrition
Left
Stayed
Count
7000
6000
5000
4000
3000
2000
1000
0
Education
Finance
Media
Healthcare
Technology
Job Role

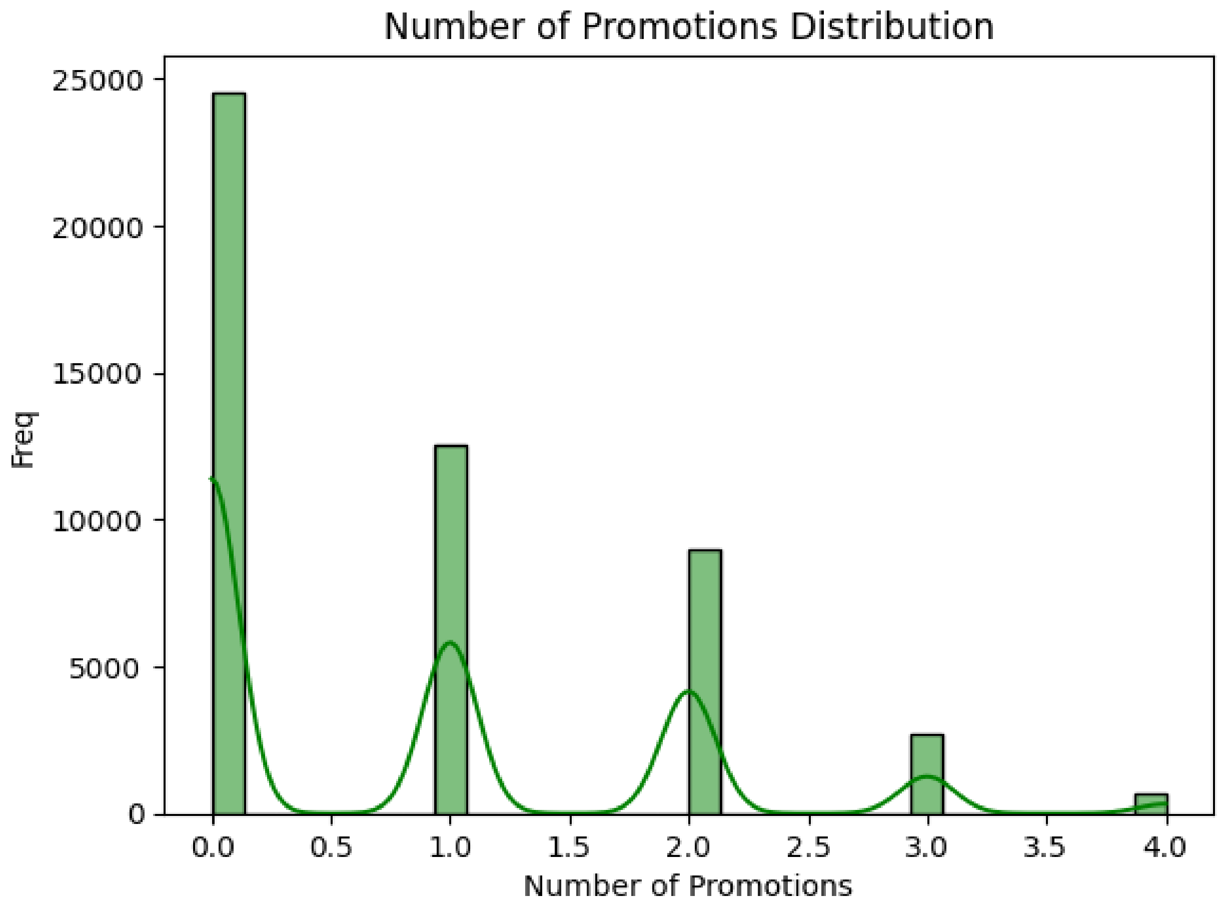

Job Satisfaction vs Attrition
Attrition
Left
Stayed
Count
14000
12000
10000
8000
6000
4000
2000
0
High
Very High
Medium
Low
Job Satisfaction

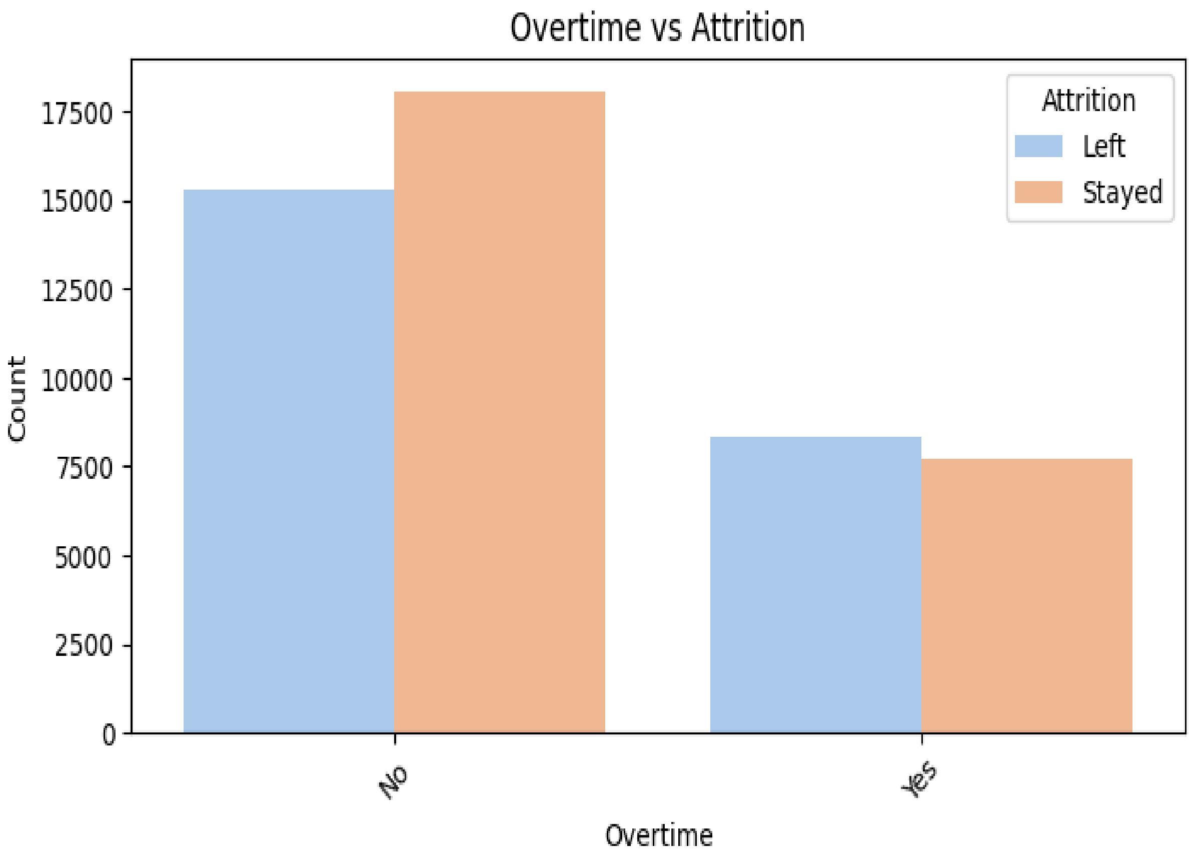

Leadership Opportunities vs Attrition
Attrition
Left
Stayed
Count
25000
20000
15000
10000
5000
0
No
Yes
Leadership Opportunities

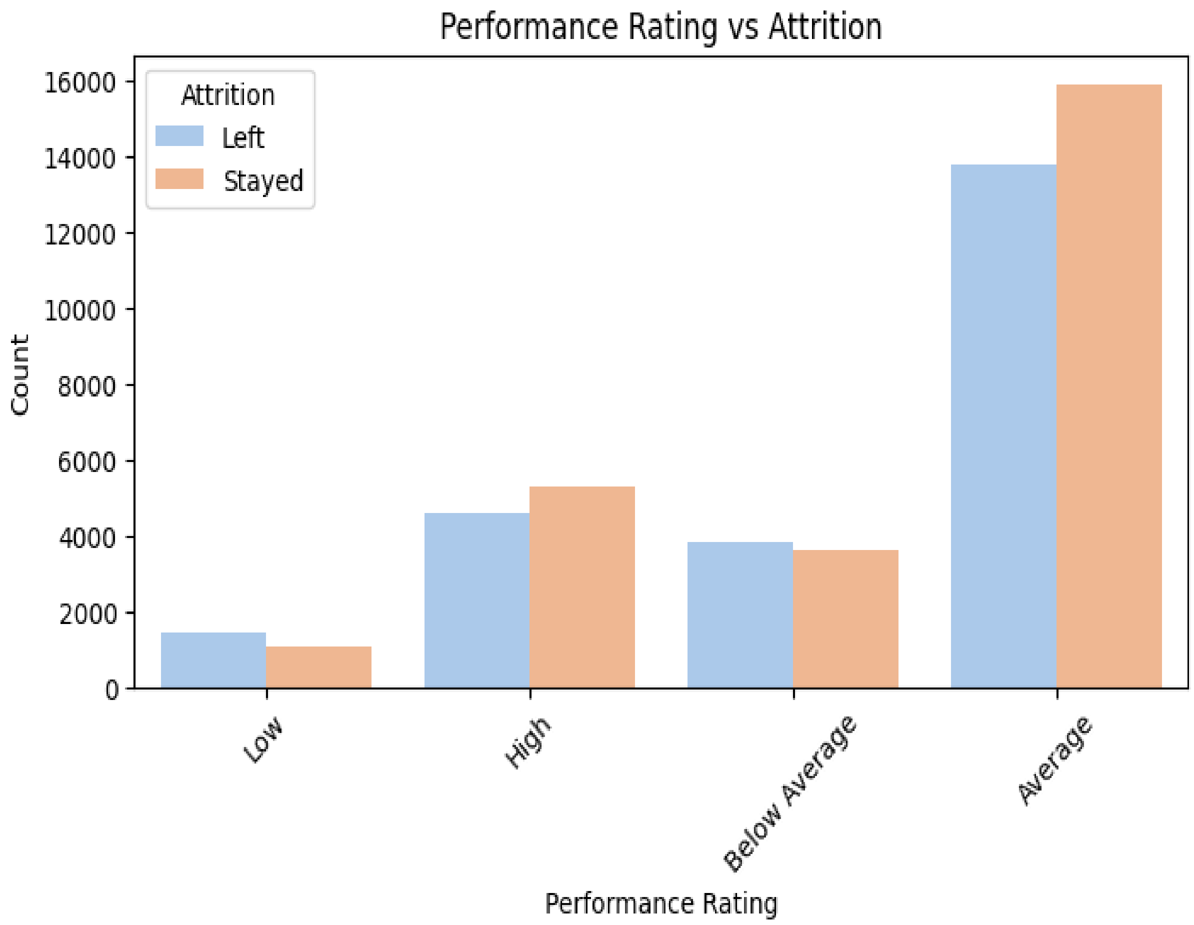

Marital Status vs Attrition
Attrition
Left
Stayed
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
Single
Married
Divorced
Marital Status

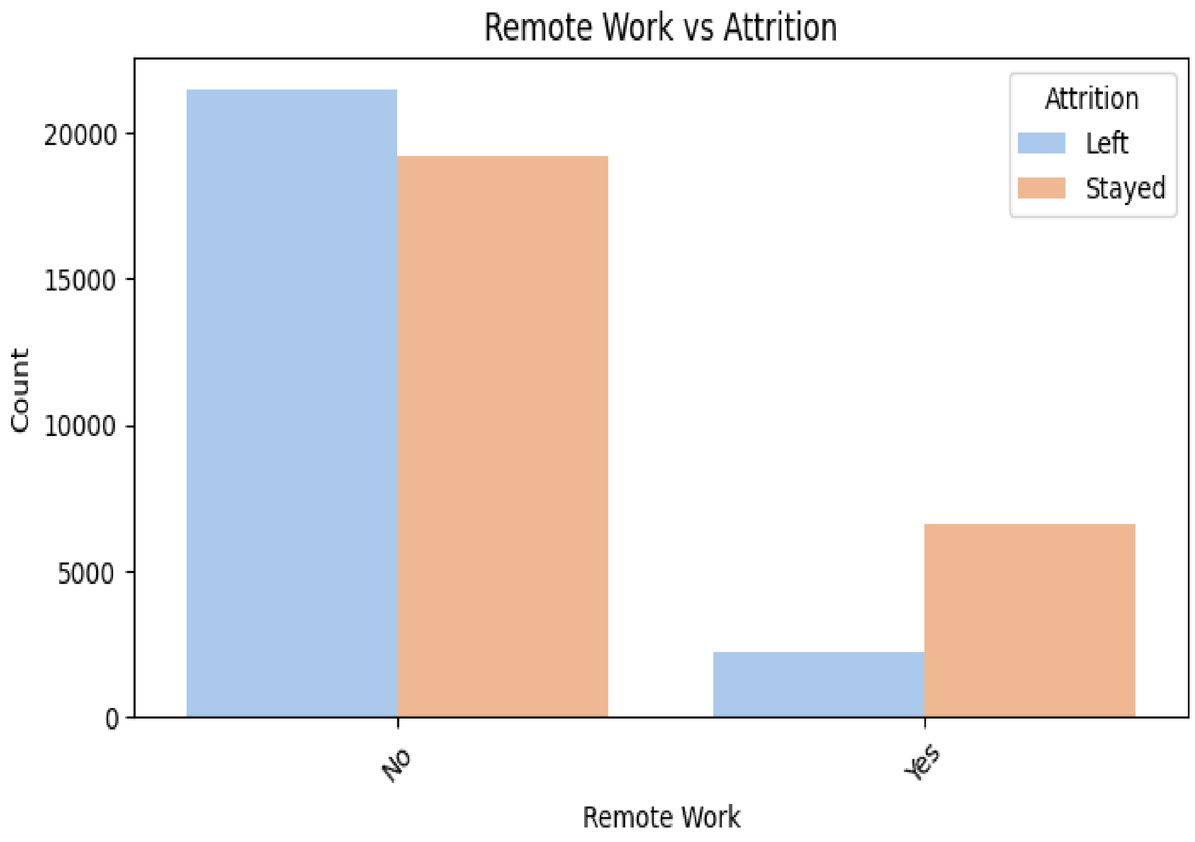

Monthly Income Distribution
Freq
0
2000
4000
6000
8000
10000
12000
14000
0
10000
20000
30000
40000
50000
Monthly Income

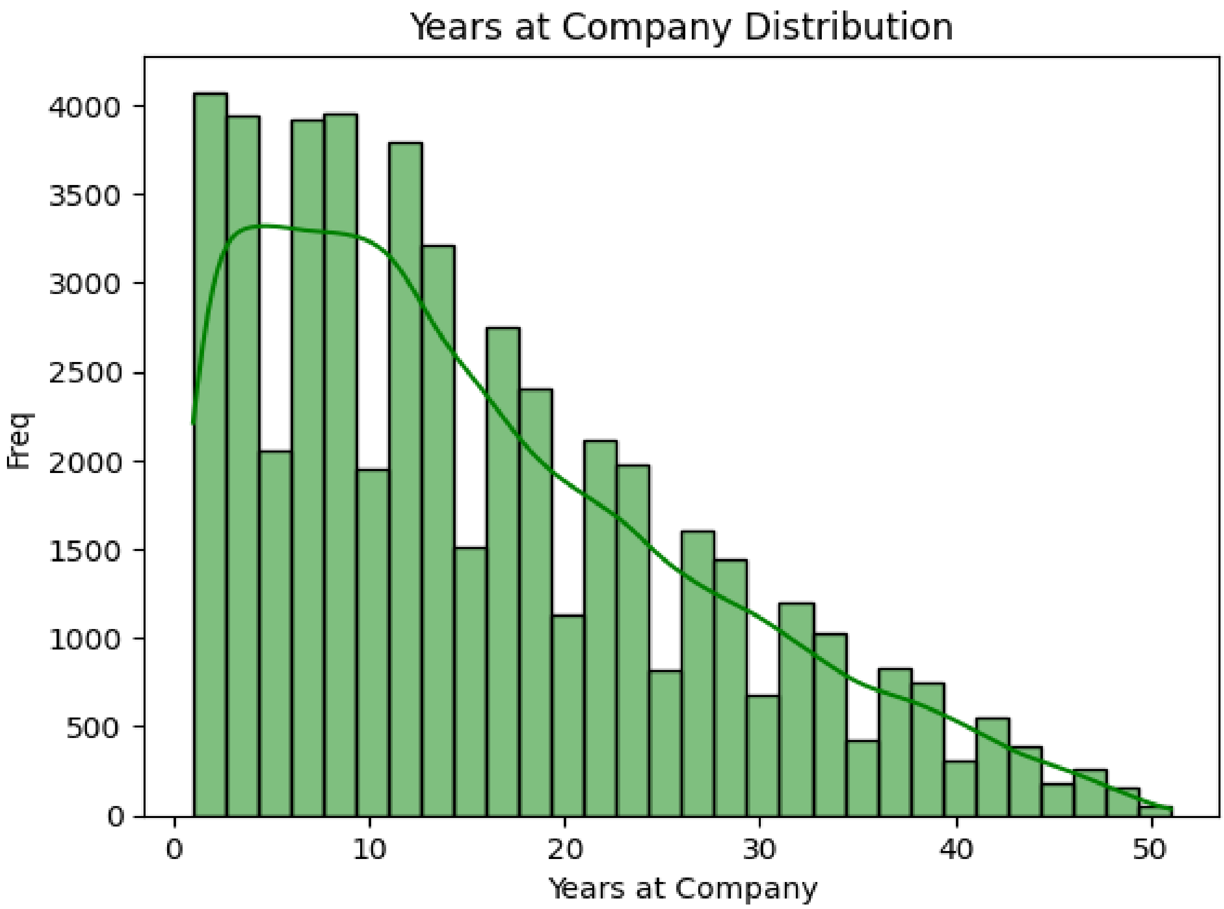

Number of Dependents Distribution
Freq
0
2000
4000
6000
8000
10000
12000
14000
16000
0
2
4
6
8
10
12
14
Number of Dependents


Number of Promotions Distribution
Freq
25000
20000
15000
10000
5000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Number of Promotions


Overtime vs Attrition
Count
17500
15000
12500
10000
7500
5000
2500
0
No
Yes
Overtime
Attrition
Left
Stayed


Performance Rating vs Attrition
Attrition
Left
Stayed
Count
0
2000
4000
6000
8000
10000
12000
14000
16000
Low
High
Below Average
Average
Performance Rating


Remote Work vs Attrition
Attrition
Left
Stayed
Count
0
5000
10000
15000
20000
No
Yes
Remote Work



7.Recommendations

- **Proactive Retention Strategies**:
  - Its main focus is on improving **job satisfaction** and **performance ratings**, as they are key predictors of employee attrition.
  - Address **work-life balance** and **overtime** issues to ensure that employees are not overburdened, which can contribute to higher turnover.
  - **Enhance company reputation** and **provide leadership opportunities**, which can improve both employee retention and satisfaction.
  - Support **single employees** and those in **entry-level positions**, as they appear to be more at risk of leaving the company.

- **Targeted Interventions**:
  - Use the model's predictions to identify at-risk employees and proactively engage with them through tailored retention efforts, such as personalized support, career development programs, and wellness initiatives.
- **Continuous Monitoring**:
  - Keep track of **company size**, **promotion opportunities**, and **education levels**, as they also play important roles in retention and should be regularly reviewed.

8.Conclusion

The final model, which uses Logistic Regression with an optimised threshold of 0.4681, has a balanced trade-off between precision and sensitivity, and it has great promise for predicting staff retention. Understanding the main causes of turnover, such as job happiness, working overtime, and corporate reputation, helps HR teams make proactive retention improvements.

By implementing these findings, the business may cultivate a more devoted workforce, lower attrition, and raise general employee satisfaction—all of which will eventually contribute to a more stable and effective work environment.