

# EMPLOYEE ATTRITION PREDICTION

## Section 1: Introduction

### Problem Statement

Employee attrition is a critical challenge for organizations, leading to higher recruitment costs, reduced productivity, and loss of institutional knowledge.

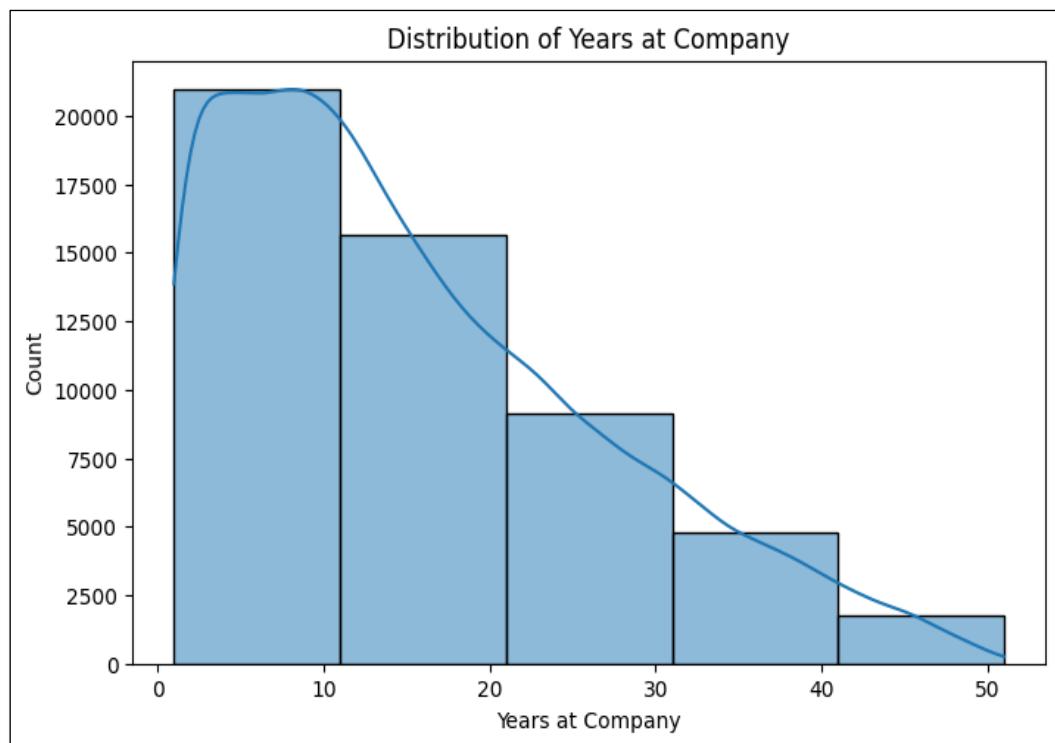
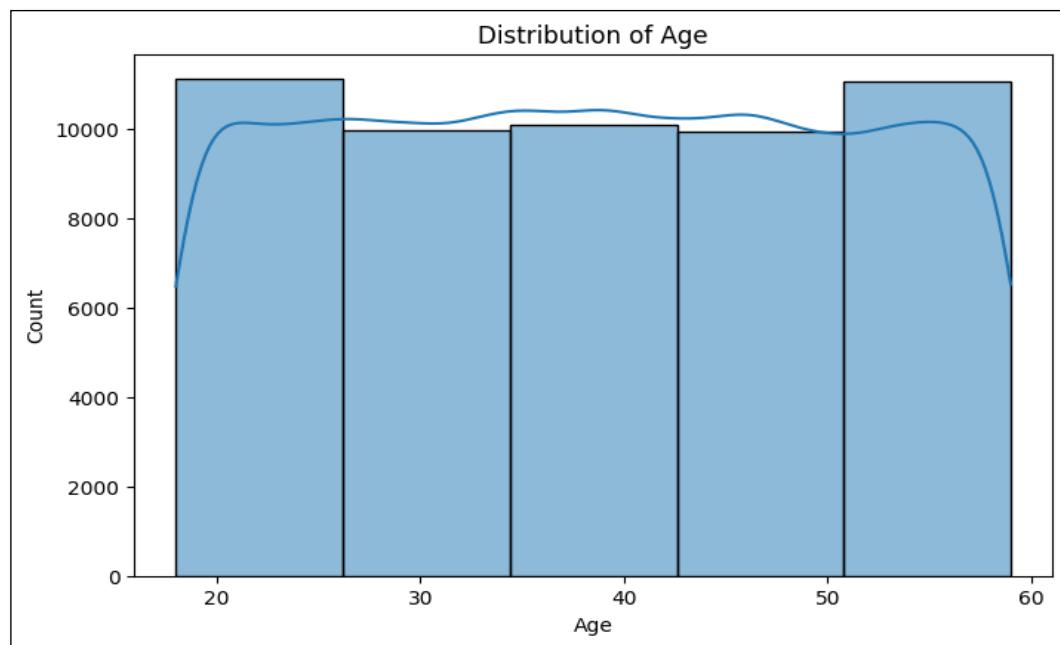
This assignment addresses the problem by developing a logistic regression model to predict whether an employee will stay or leave, enabling HR teams to design proactive retention strategies.

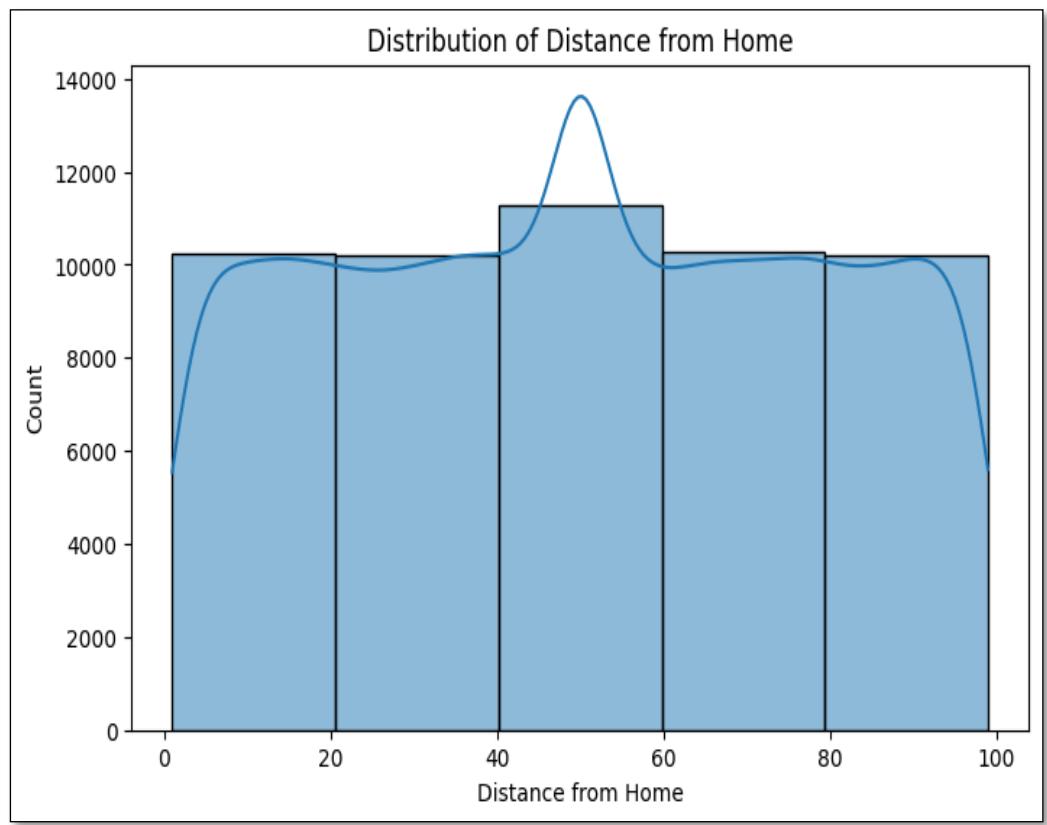
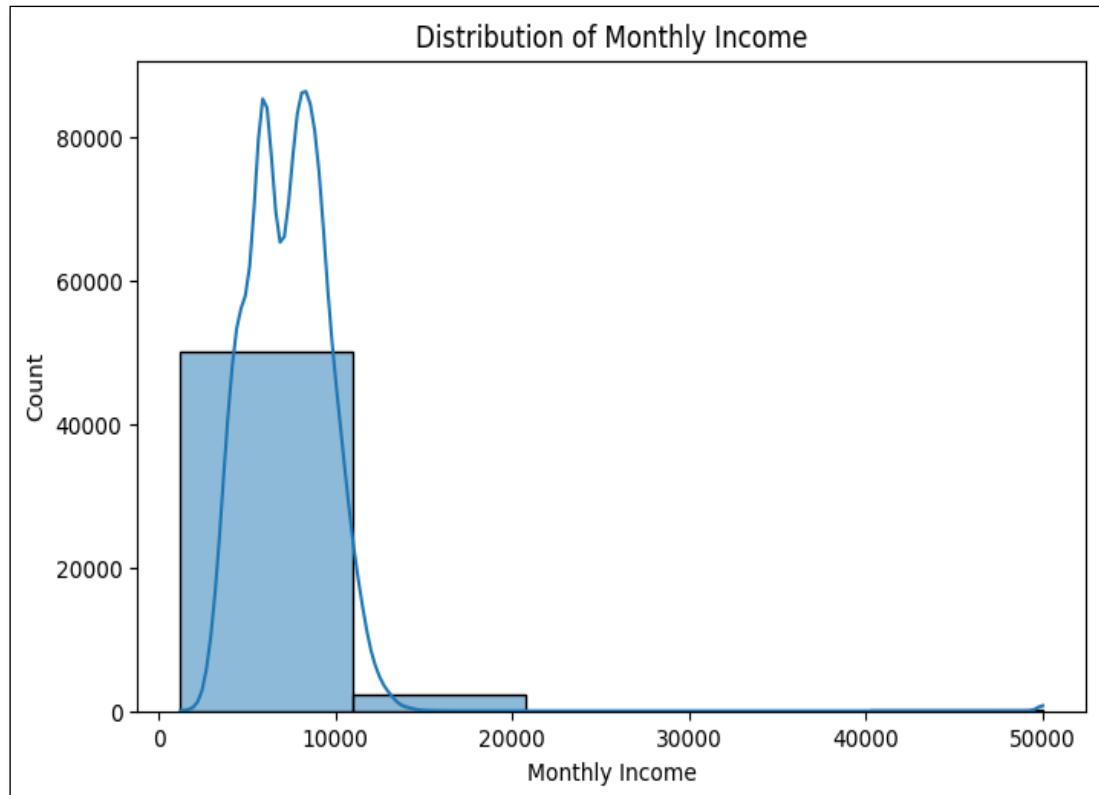
### Methodology & Techniques Used

- Exploratory Data Analysis (EDA): univariate, bivariate, multivariate, and correlation analysis.
- Data Preparation: one-hot encoding, Boolean conversion, feature selection, train/validation split.
- Modelling: logistic regression using statsmodels, ROC curve analysis, cutoff selection.
- Evaluation: confusion matrices, accuracy, sensitivity, specificity, precision, recall.
- Insights: interpretation of regression coefficients to identify retention and attrition drivers.

## Section 2: Exploratory Data Analysis (EDA)

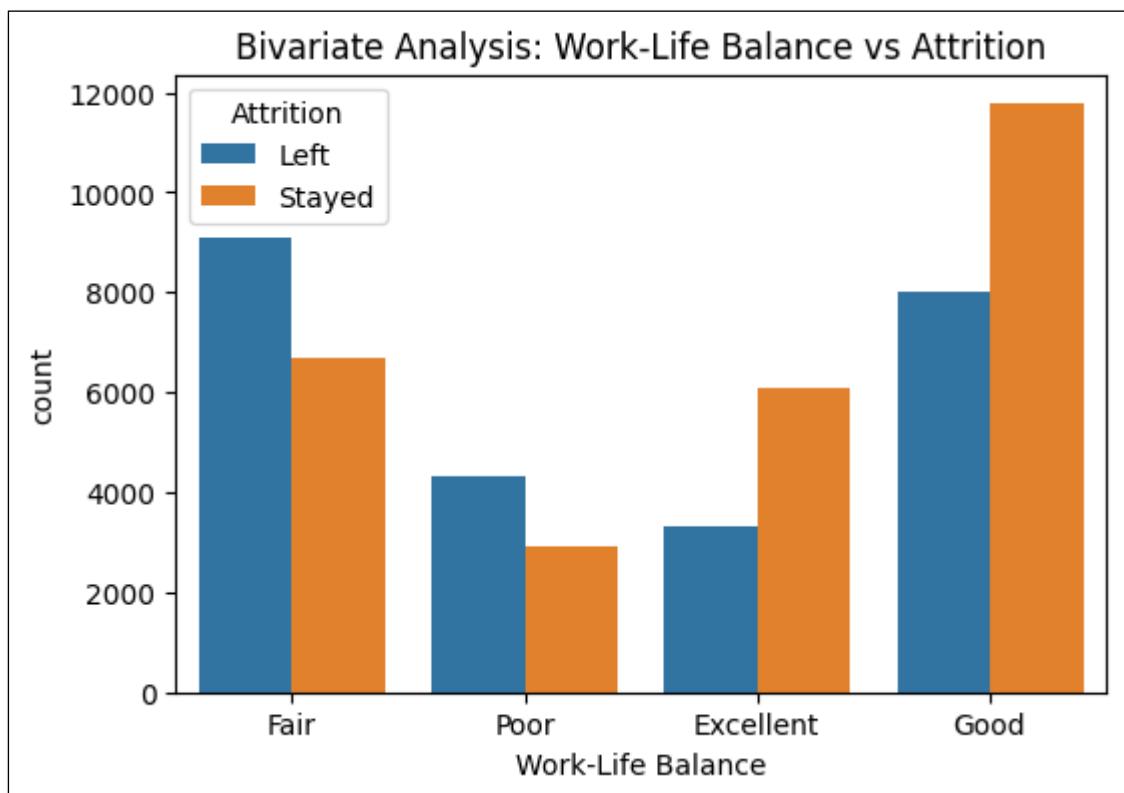
### 2.1 Univariate Analysis

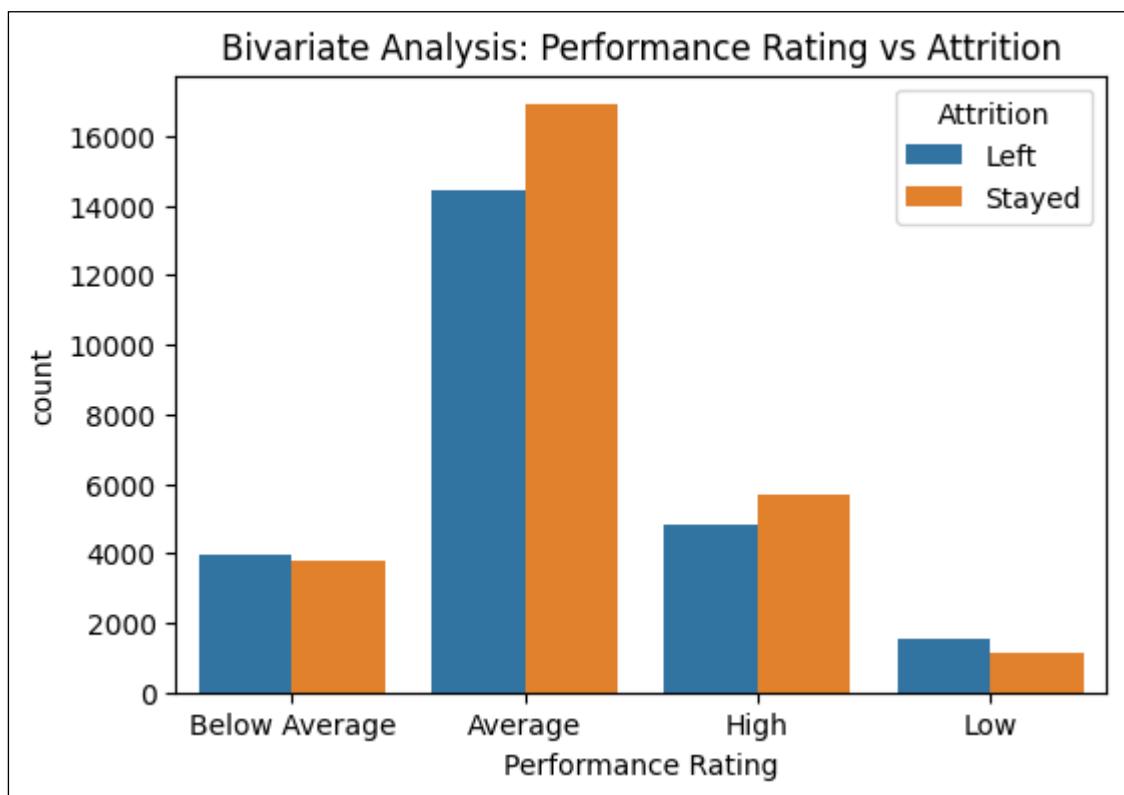
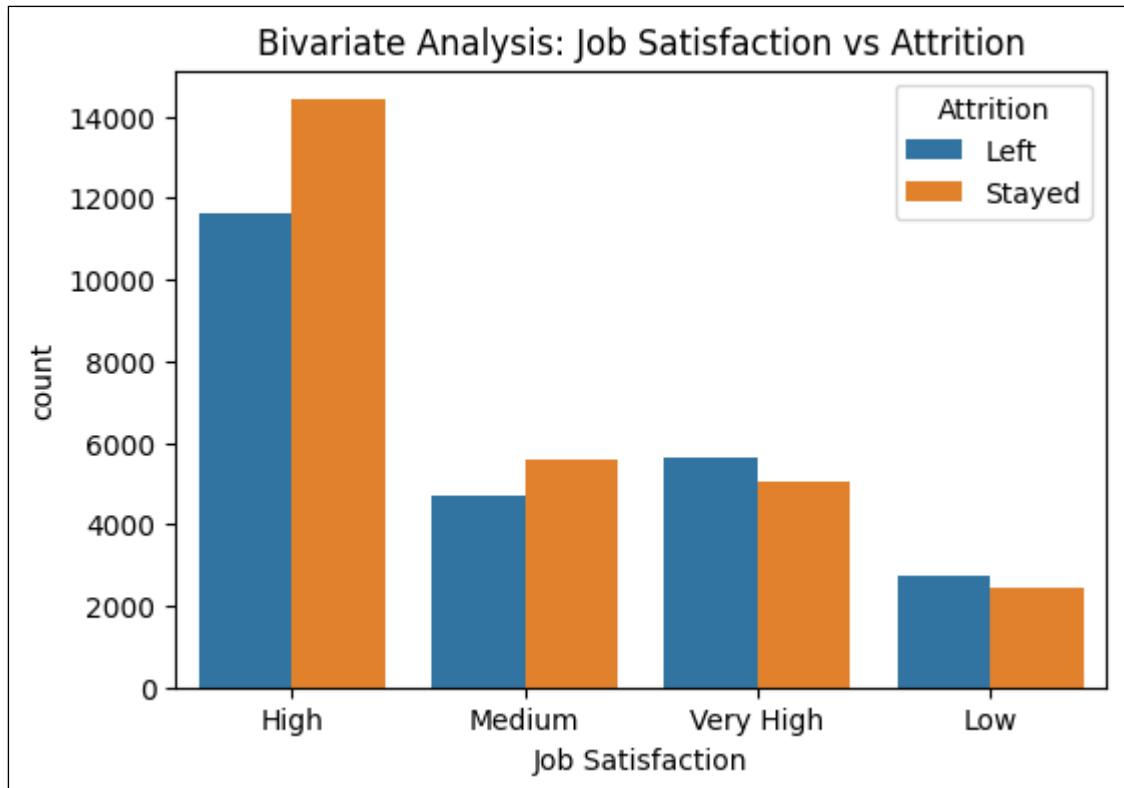


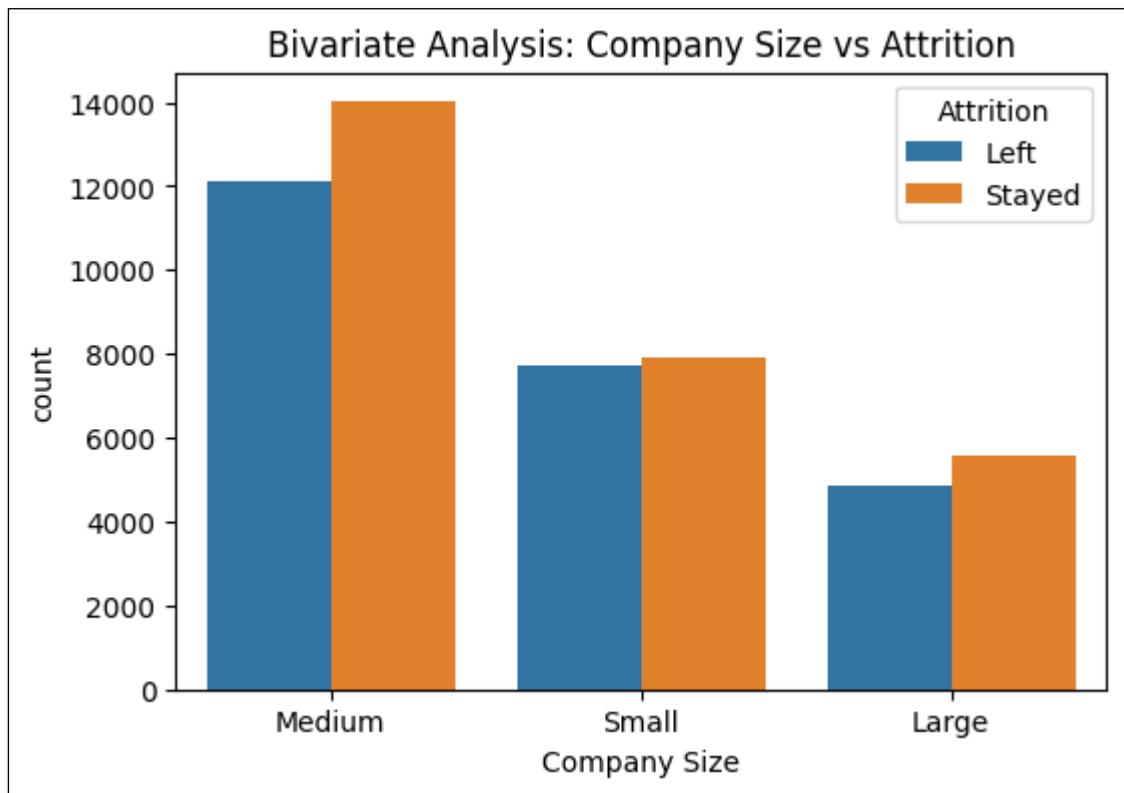
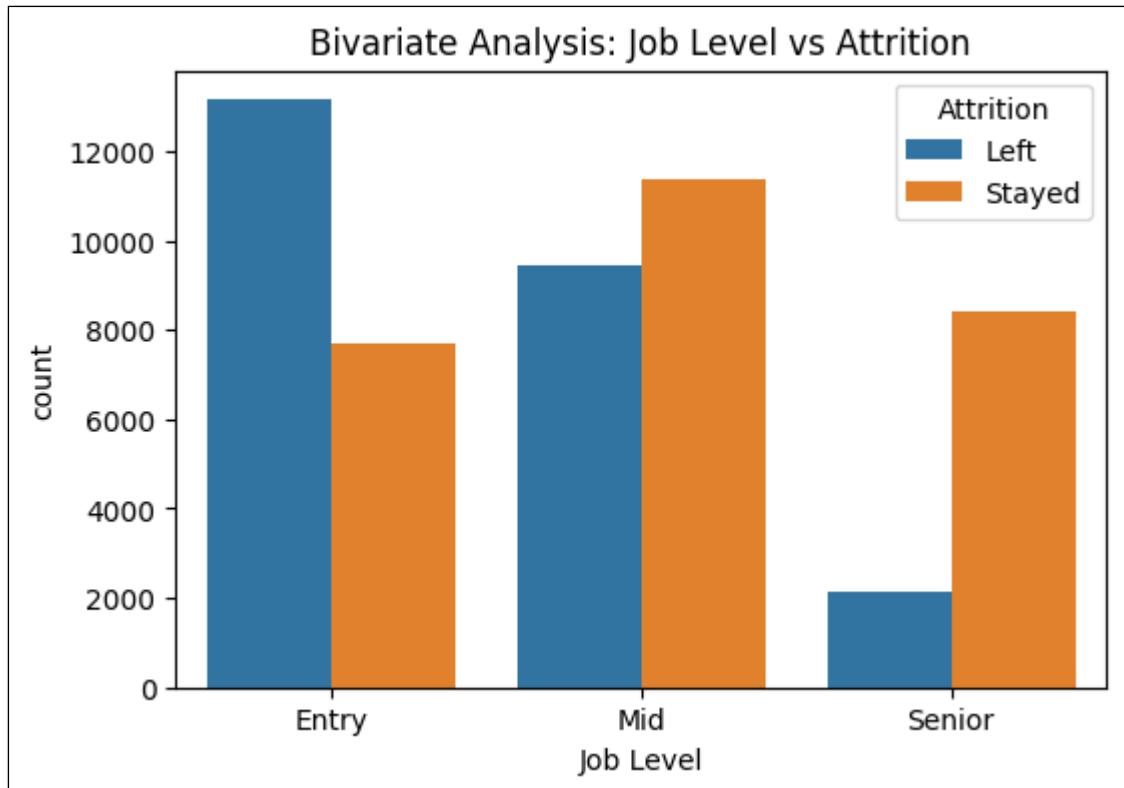


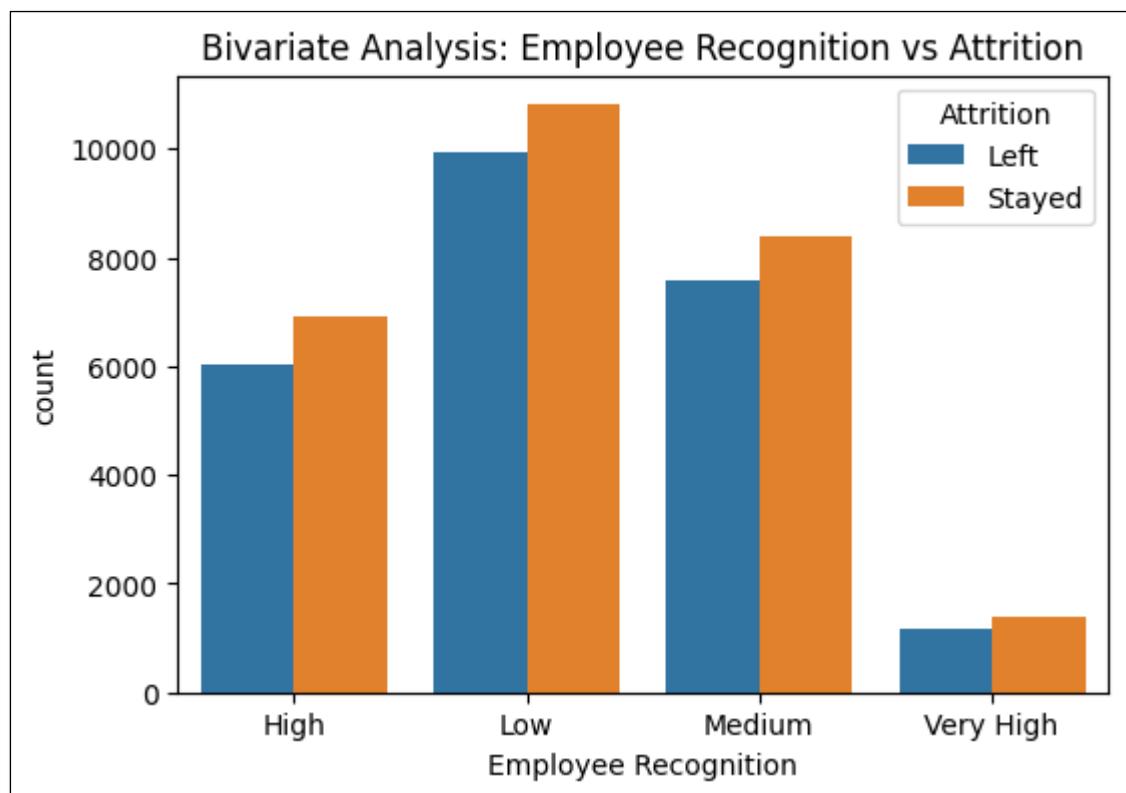
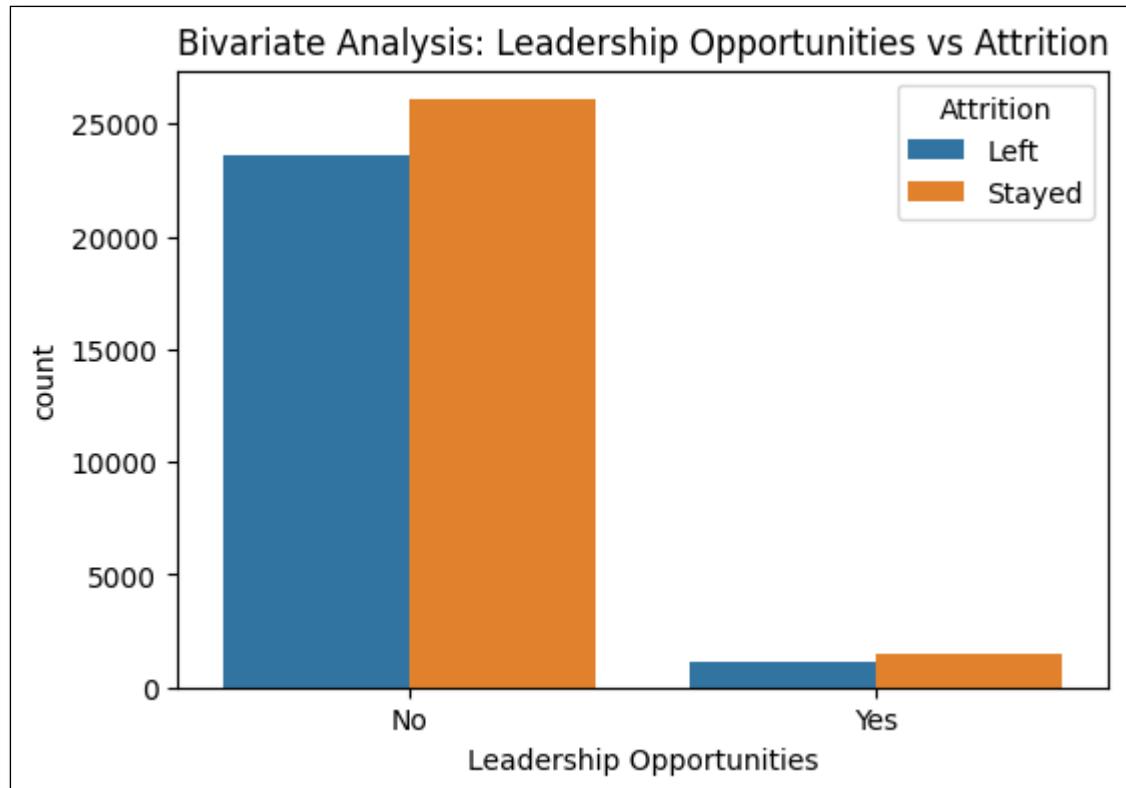
## 2.2 Bivariate Analysis

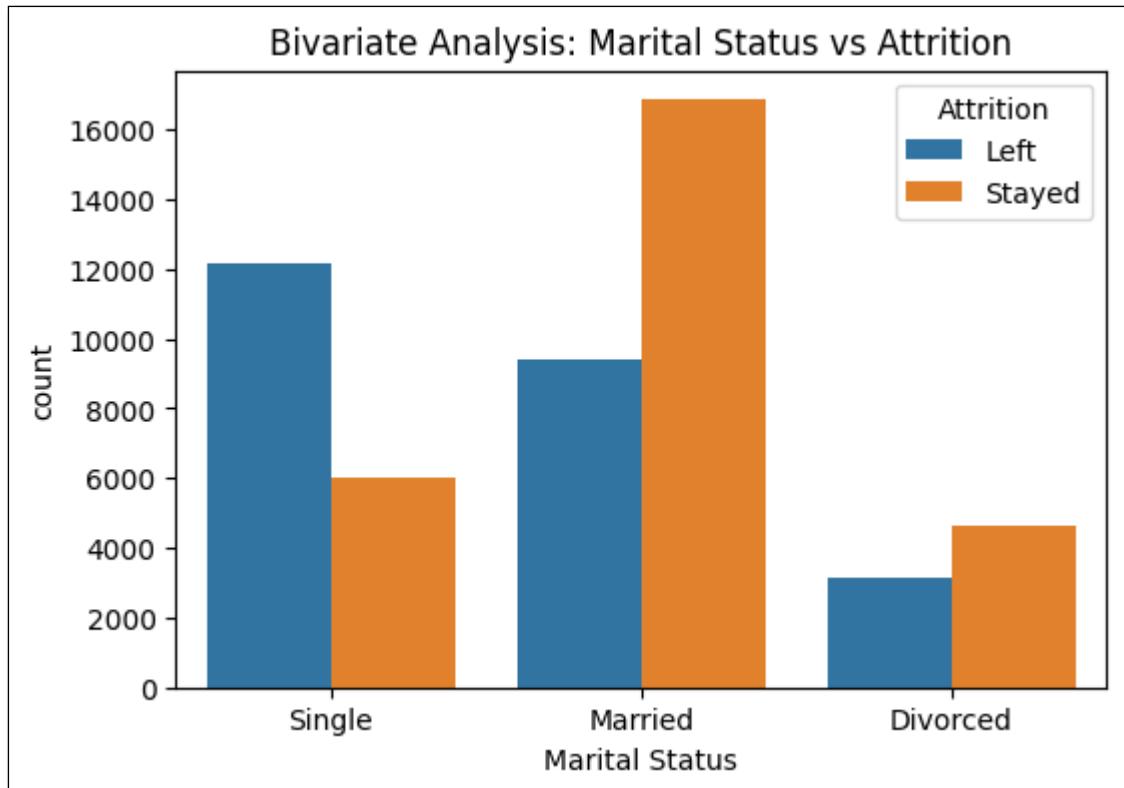
- **Purpose:** Explore relationships between predictors and the target (Stayed vs Attrition).
- **Remote Work:** Employees working remotely show higher retention rates.
- **Job Level:** Attrition is higher among junior employees; mid and senior levels show stronger retention.
- **Education Level:** PhD holders are more likely to stay.
- **Overtime:** Employees working overtime are more likely to leave.
- **Marital Status:** Single employees show higher attrition compared to married ones.
- **Company Reputation:** Poor reputation correlates with higher attrition.



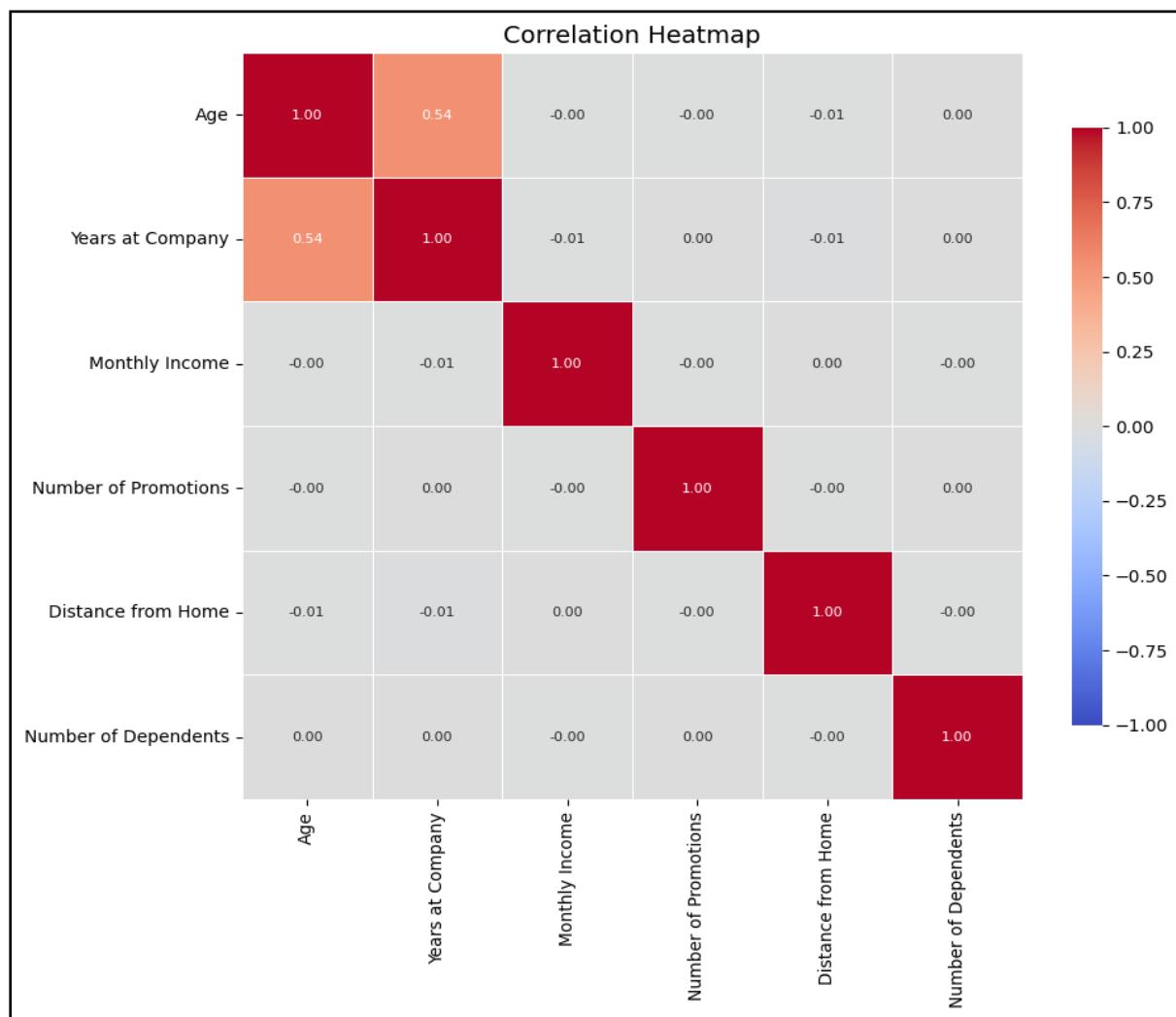








## 2.3 Multivariate Analysis



### Key Outcomes

- **Age & Years at Company** show a moderate positive correlation (**0.54**), indicating that older employees tend to have longer tenure.
- All other numeric features — **Monthly Income, Number of Promotions, Distance from Home, and Number of Dependents** — show **very weak or no correlation** with each other (values close to 0).

## 4. Data Preparation

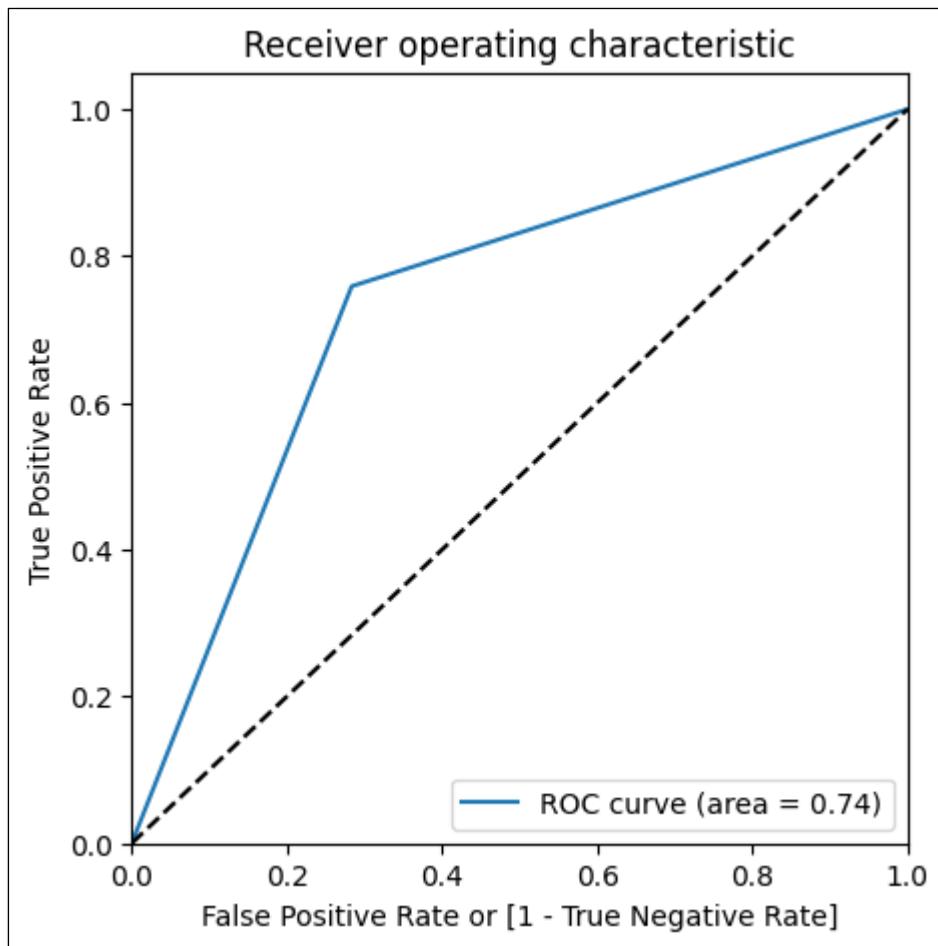
- Steps Taken
  - One-hot encoding for categorical variable
  - Converted boolean columns to integers.
  - Added constant for statsmodels regression.
  - Train/validation split.
- Feature Selection:
  - Gender\_Male
  - Work-Life Balance\_Fair
  - Work-Life Balance\_Poor
  - Job Satisfaction\_Low
  - Job Satisfaction\_Very High
  - Performance Rating\_Below Average
  - Performance Rating\_Low
  - Overtime\_Yes
  - Education Level\_PhD
  - Marital Status\_Single
  - Job Level\_Mid
  - Job Level\_Senior
  - Remote Work\_Yes
  - Company Reputation\_Fair
  - Company Reputation\_Poor

## 5. Modelling

- Regression Summary

Generalized Linear Model Regression Results							
Dep. Variable:	Stayed	No. Observations:	52227				
Model:	GLM	Df Residuals:	52211				
Model Family:	Binomial	Df Model:	15				
Link Function:	Logit	Scale:	1.0000				
Method:	IRLS	Log-Likelihood:	-26405.				
Date:	Sun, 23 Nov 2025	Deviance:	52810.				
Time:	16:25:53	Pearson chi2:	4.89e+04				
No. Iterations:	5	Pseudo R-squ. (CS):	0.3109				
Covariance Type:	nonrobust						
		coef	std err	z	P> z	[0.025	0.975]
	const	0.2366	0.028	8.571	0.000	0.183	0.291
	Gender_Male	0.5909	0.022	27.327	0.000	0.549	0.633
	Work-Life Balance_Fair	-1.0674	0.025	-43.545	0.000	-1.115	-1.019
	Work-Life Balance_Poor	-1.2243	0.033	-37.397	0.000	-1.288	-1.160
	Job Satisfaction_Low	-0.4542	0.036	-12.525	0.000	-0.525	-0.383
	Job Satisfaction_Very High	-0.4738	0.027	-17.622	0.000	-0.526	-0.421
	Performance Rating_Below Average	-0.3205	0.030	-10.658	0.000	-0.379	-0.262
	Performance Rating_Low	-0.5769	0.049	-11.701	0.000	-0.674	-0.480
	Overtime_Yes	-0.3641	0.023	-15.964	0.000	-0.409	-0.319
	Education Level_PhD	1.5191	0.054	27.914	0.000	1.412	1.626
	Marital Status_Single	-1.7098	0.024	-71.208	0.000	-1.757	-1.663
	Job Level_Mid	0.9562	0.024	40.582	0.000	0.910	1.002
	Job Level_Senior	2.5084	0.034	74.339	0.000	2.442	2.575
	Remote Work_Yes	1.7126	0.030	56.448	0.000	1.653	1.772
	Company Reputation_Fair	-0.4897	0.028	-17.618	0.000	-0.544	-0.435
	Company Reputation_Poor	-0.7232	0.028	-26.133	0.000	-0.777	-0.669

- **ROC Curve:**



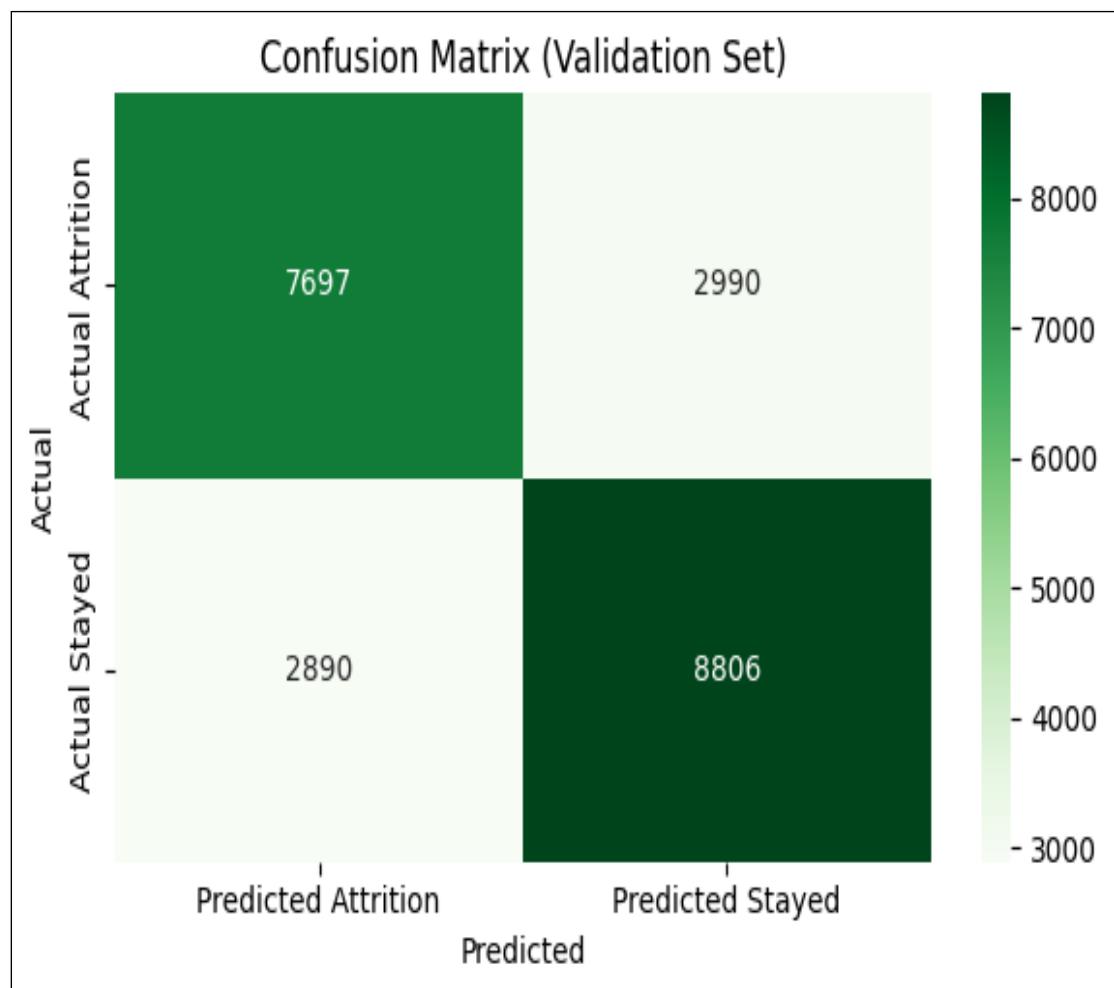
- **Cutoff Selection:** Optimal threshold chosen at **0.5**

## 6. Results

- Performance Metrics Table

Metric	Value
Accuracy	0.7373
Sensitivity	0.7529
Specificity	0.7202
Precision	0.7465
Recall	0.7529

- Confusion Matrix



- Final Prediction Data Frame:

	<b>Actual</b>	<b>Predicted</b>	<b>final_prediction</b>
0	True	0.989411	1
1	True	0.696075	1
2	False	0.208089	0
3	False	0.413788	0
4	False	0.064596	0

## 7. Insights

- **Retention Drivers (positive coefficients → Stayed = 1):**
  - Male employees, PhD holders, Mid/Senior job levels, Remote workers.
- **Attrition Drivers (negative coefficients → Attrition = 1):**
  - Poor work-life balance, low job satisfaction, single marital status, low performance ratings.

## 8. Conclusion

- The logistic regression model achieved ~74% accuracy with balanced sensitivity and specificity.
- While 90% accuracy is unlikely due to the complexity of attrition behaviours, the model provides actionable insights for HR.
- The model can be used as a screening tool to identify at-risk employees and guide retention strategies.