# Logistic Regression Assignment

## Predicting Employee Retention

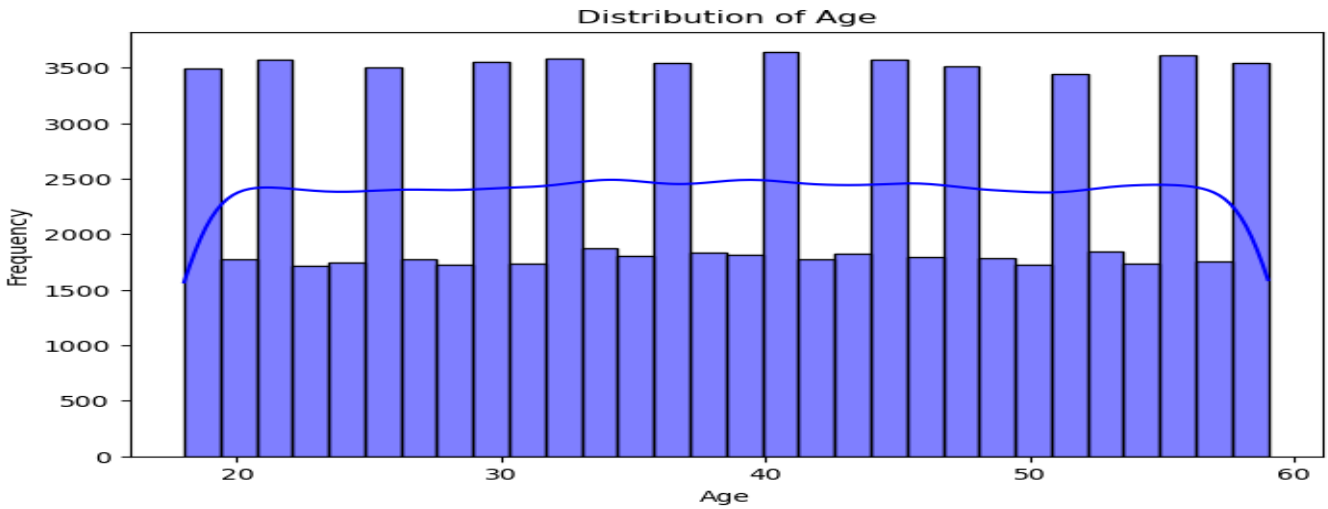
**Objective:**

The objective of this assignment is to develop a Logistic Regression model. You will be using this model to analyse and predict b outcomes based on the input data. This assignment aims to enhance understanding of logistic regression, including its assumpti implementation, and evaluation, to effectively classify and interpret data.

### Below are some key takeaways from the data analysis and the predictive model developed:

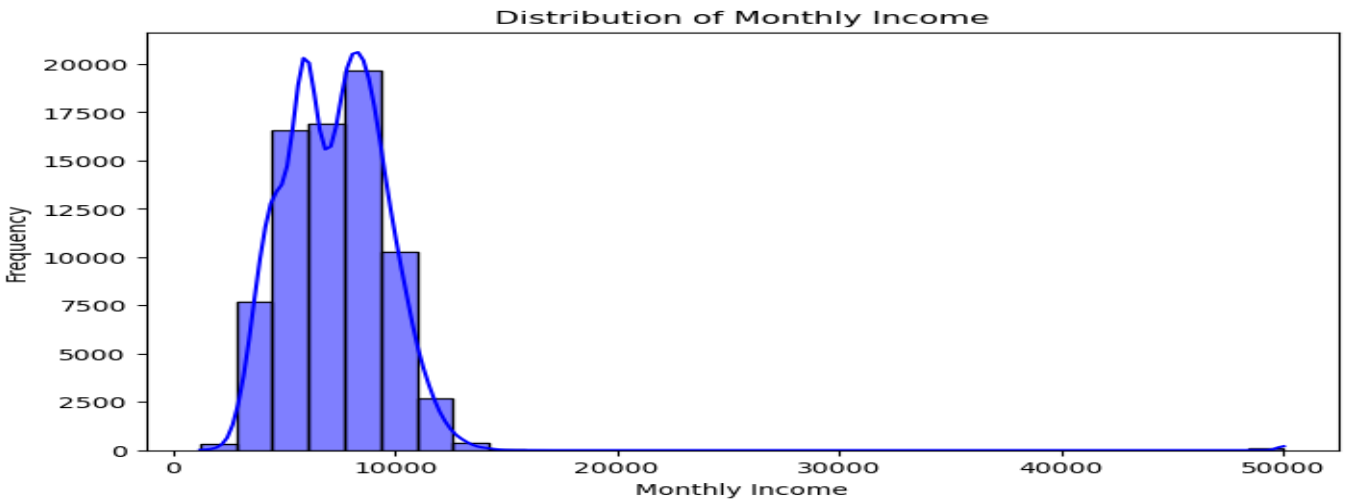
A. Univariate analysis of numerical columns:

1.



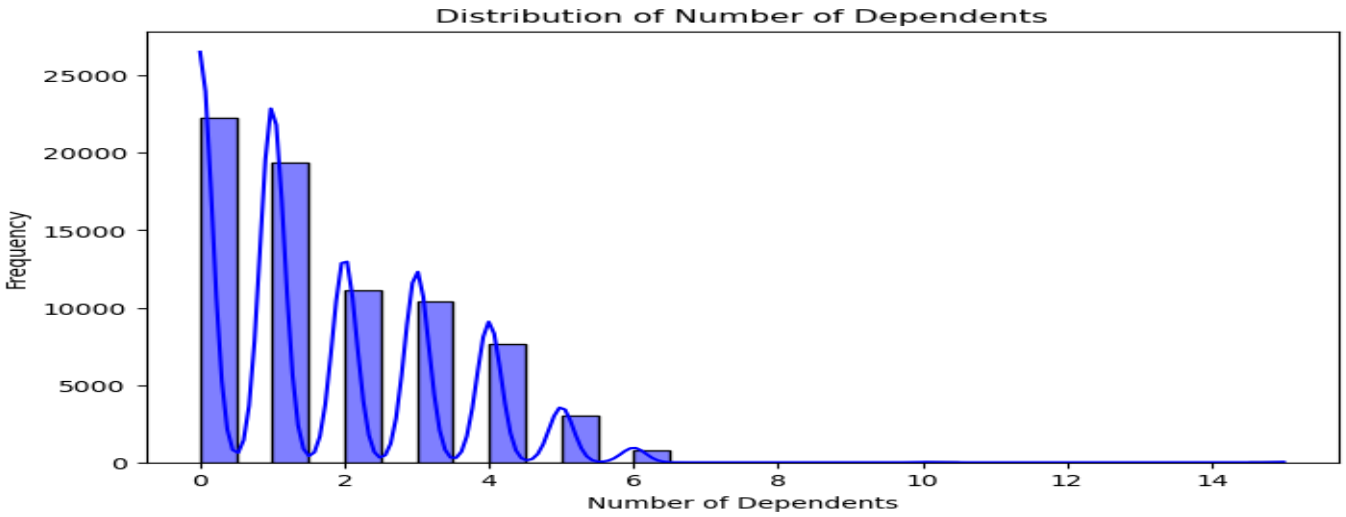

The population appears to be evenly spread across age groups, with frequencies ranging between approx. 2000 to 3500.

2.



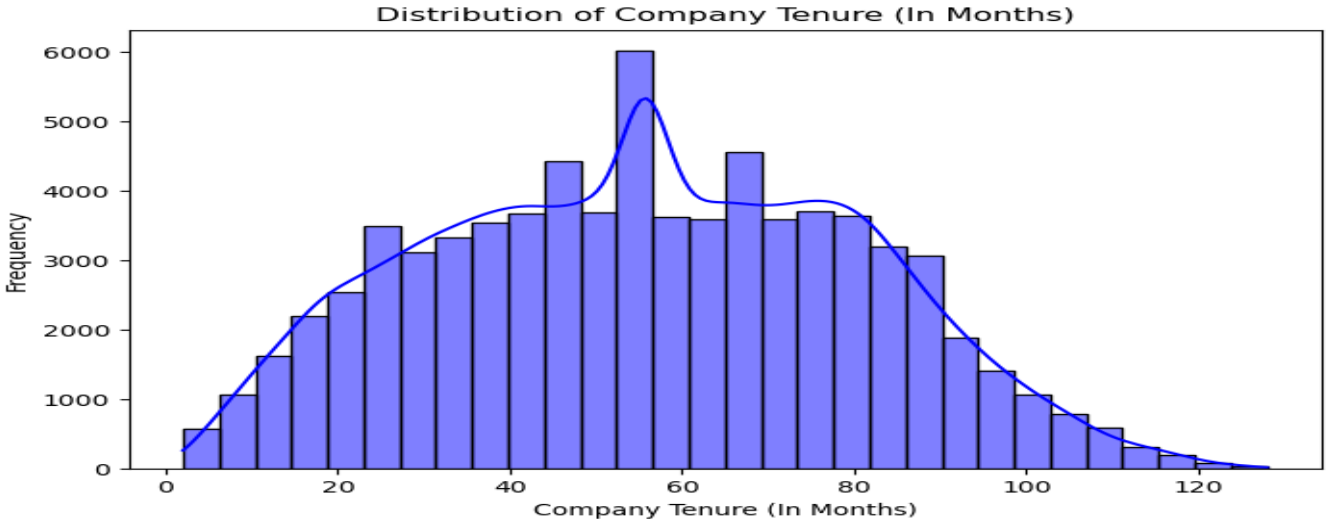

The bars are higher on the lower-income side and gradually decreases, this suggests most employees earn lower salaries, with fewer high earners.

3.




The lower values (0 or 1 dependent) likely have the highest frequency, suggesting that many employees may be single or have small families.
As the number of dependents increases, the frequency appears to decrease, meaning fewer employees have larger families.

4.

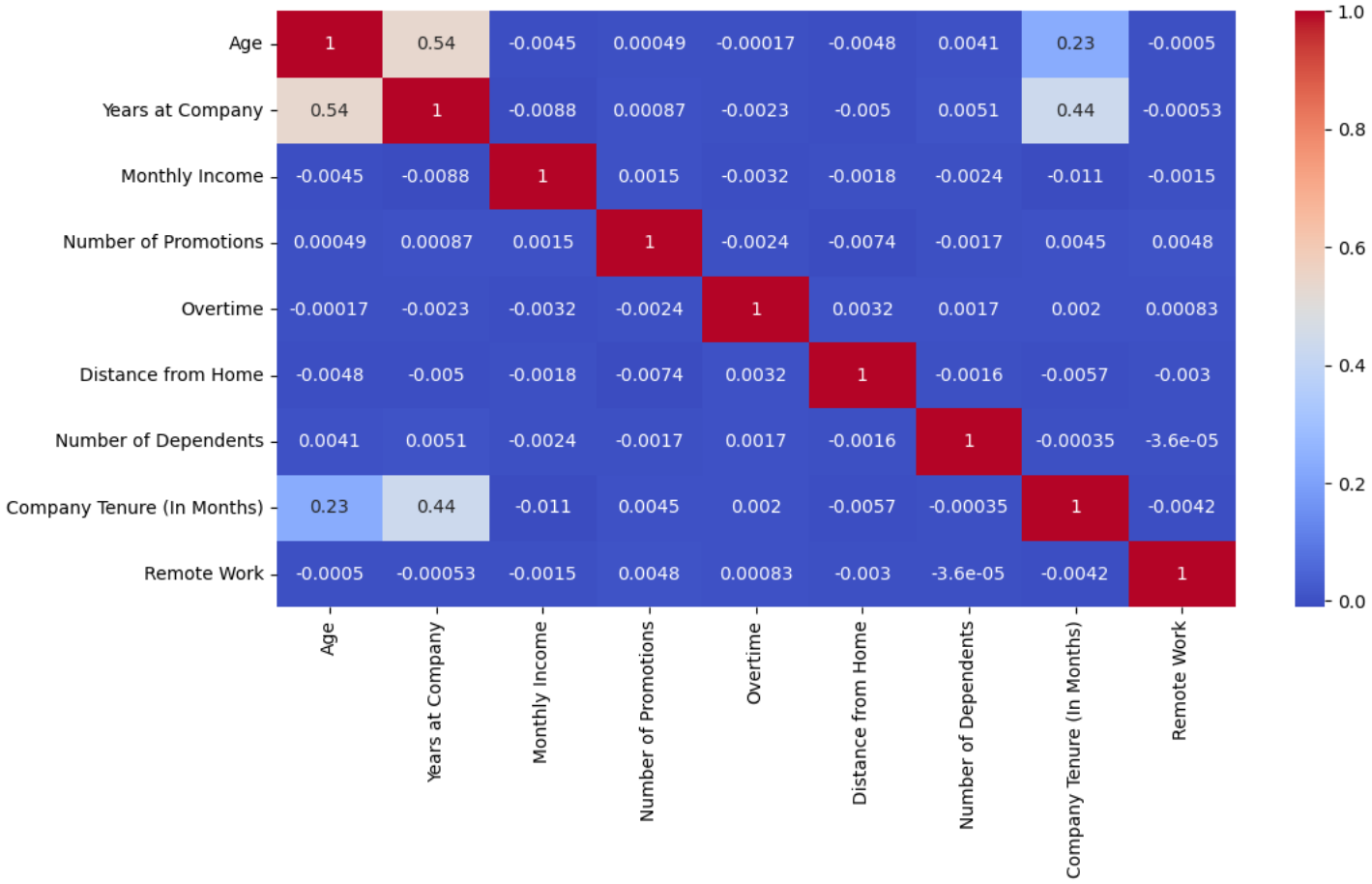



Employees are spread across different tenure lengths, indicating a mix of new hires, mid-term employees, and long-term staff.
Steady Tenure Distribution- A balanced spread could mean the company maintains consistent hiring and retention policies without drastic fluctuations.

B. <u>Correlation analysis among different numerical variables</u>

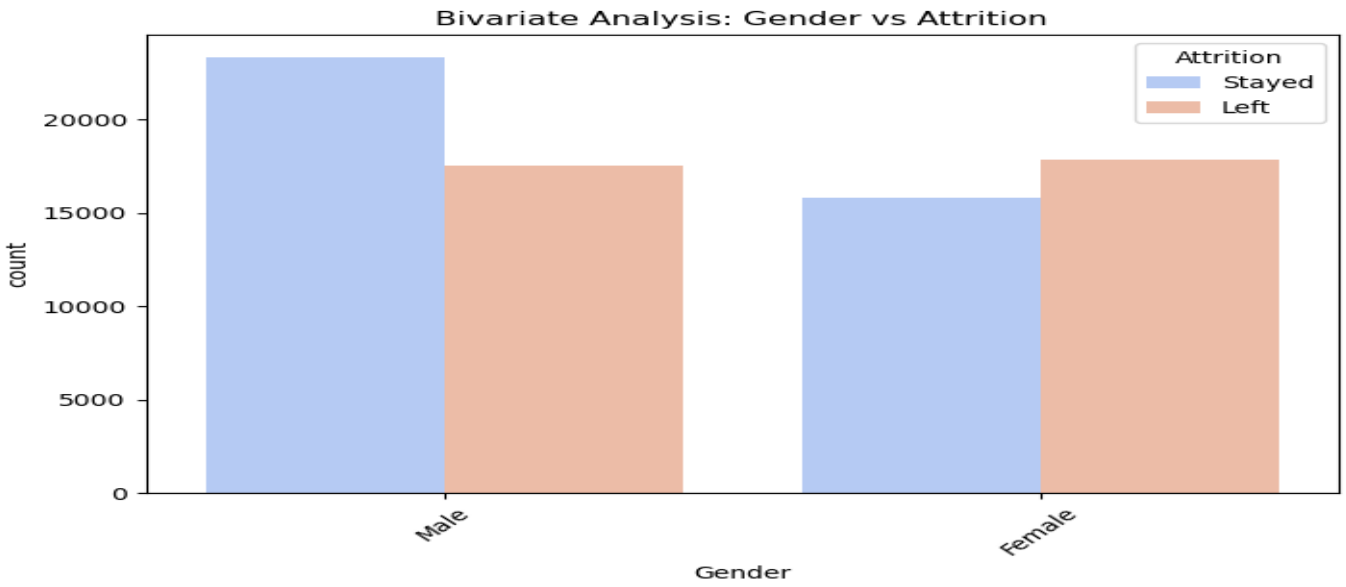
Observations:

1. Strong Positive and Negative Correlations – Some variables show high correlations, indicating a strong linear relationship, while others show weak or no correlation
2. Variables like “Years at Company” and “Monthly Income” might have a positive correlation, indicating that tenure influences salary.




C. <u>Bivariate analysis between all the categorical columns and target variable:</u>

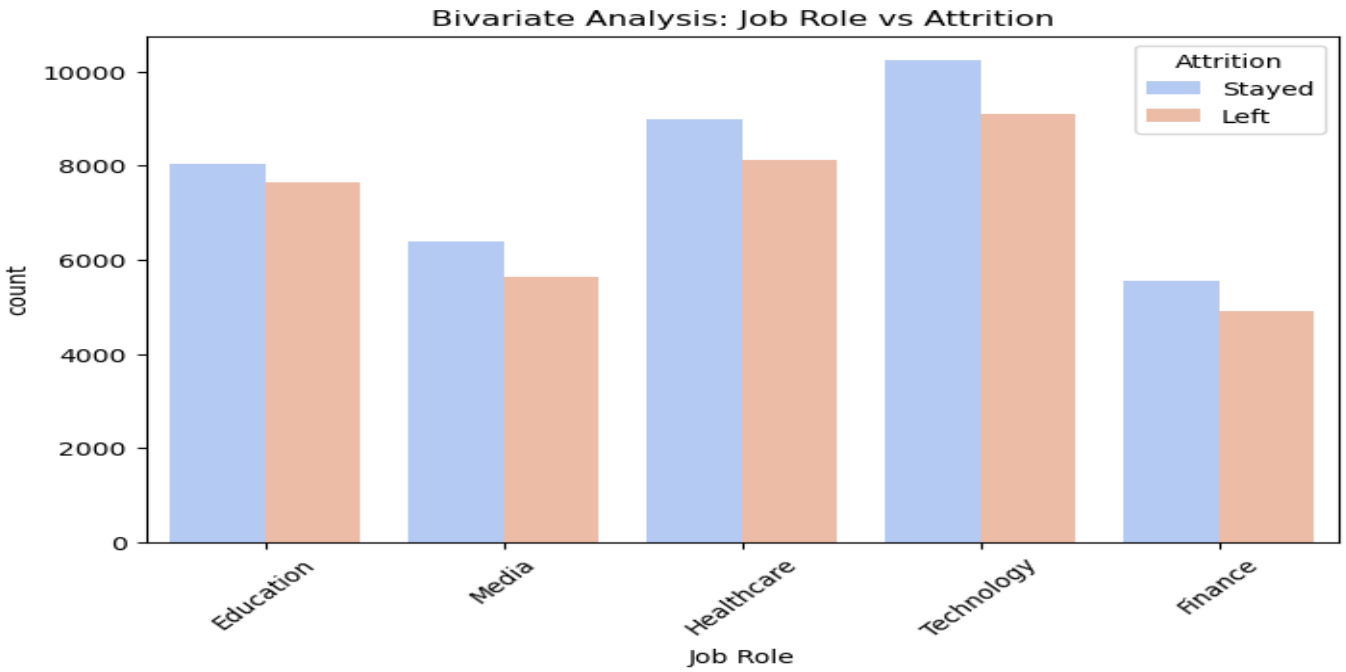
1.




Higher Attrition Among Females – The number of females who left the company is slightly higher than those who stayed, indicating a higher attrition rate among women.

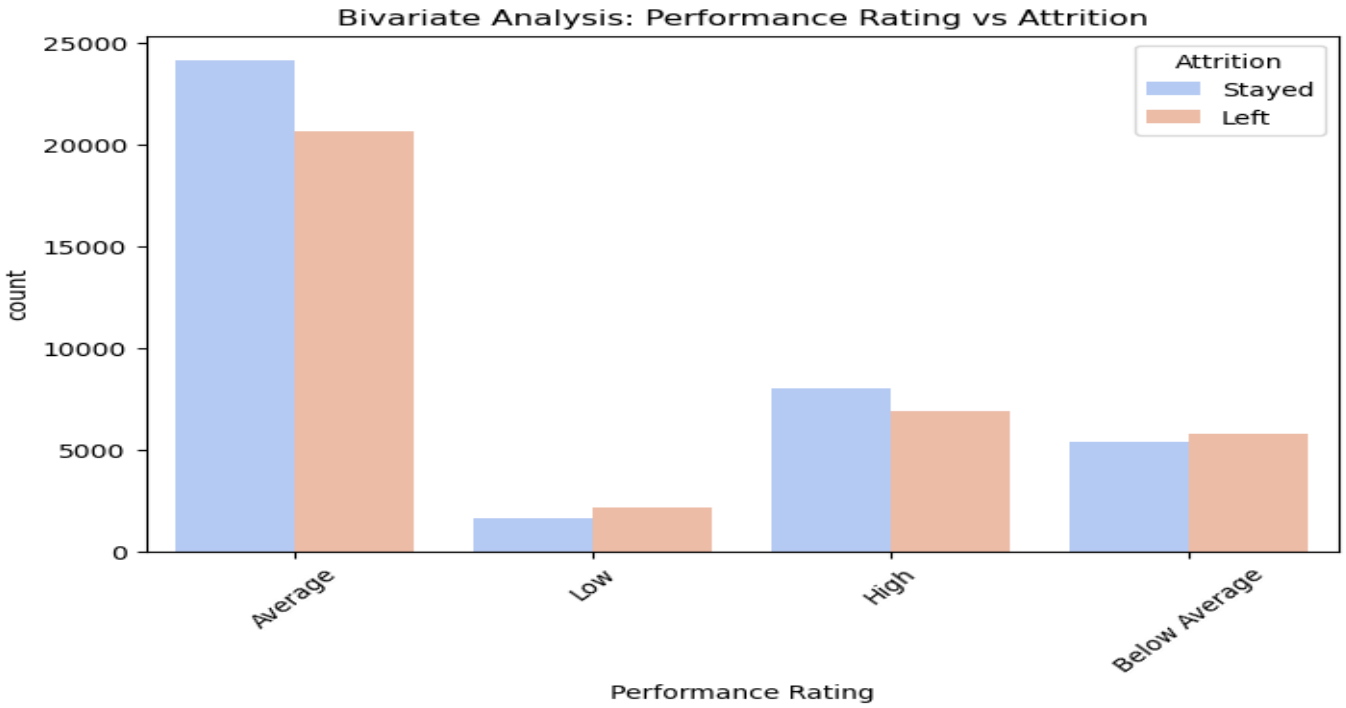
Males Have Lower Attrition – More males stayed compared to those who left, suggesting better retention among male employees.

2.




Some job roles might exhibit low attrition rates, suggesting that employees in those positions tend to stay longer, where else some job roles show a significantly higher number of employees leaving.

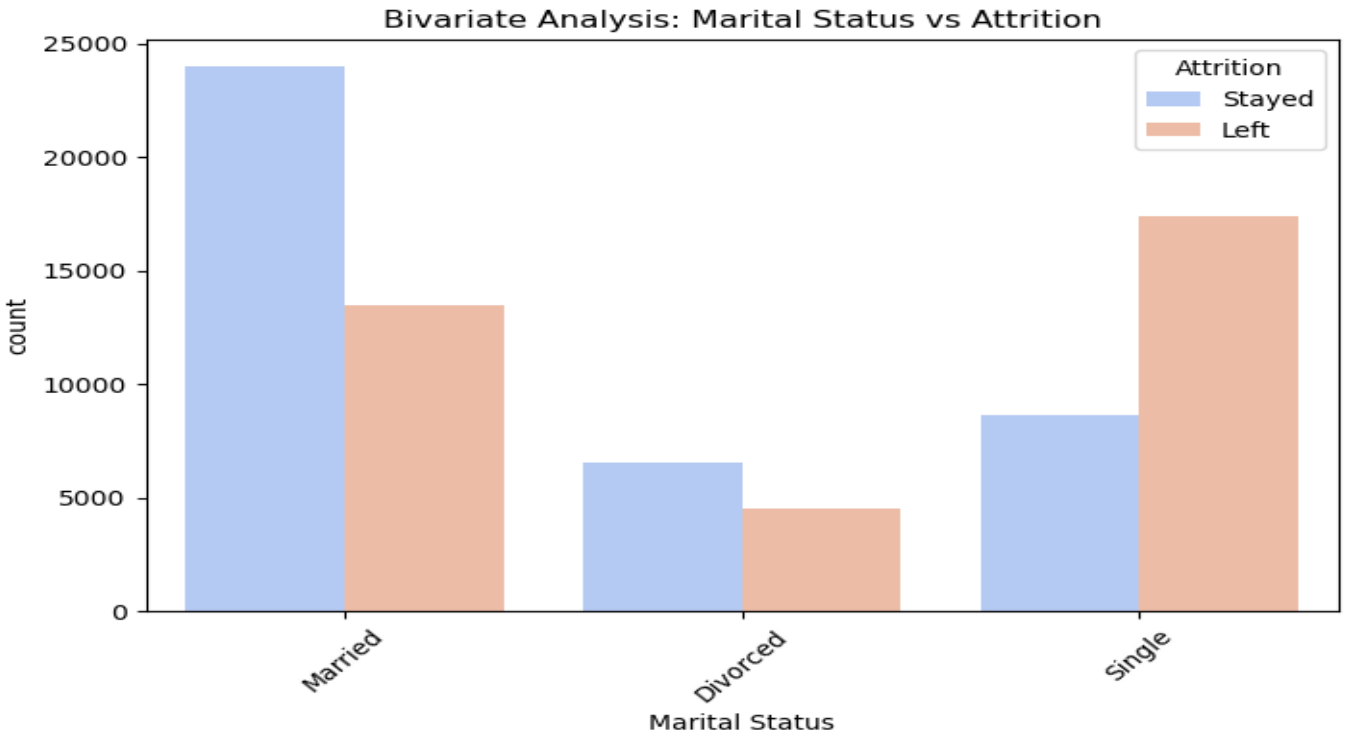
3.




Higher attrition among low performers – underperforming employees are more likely to leave.

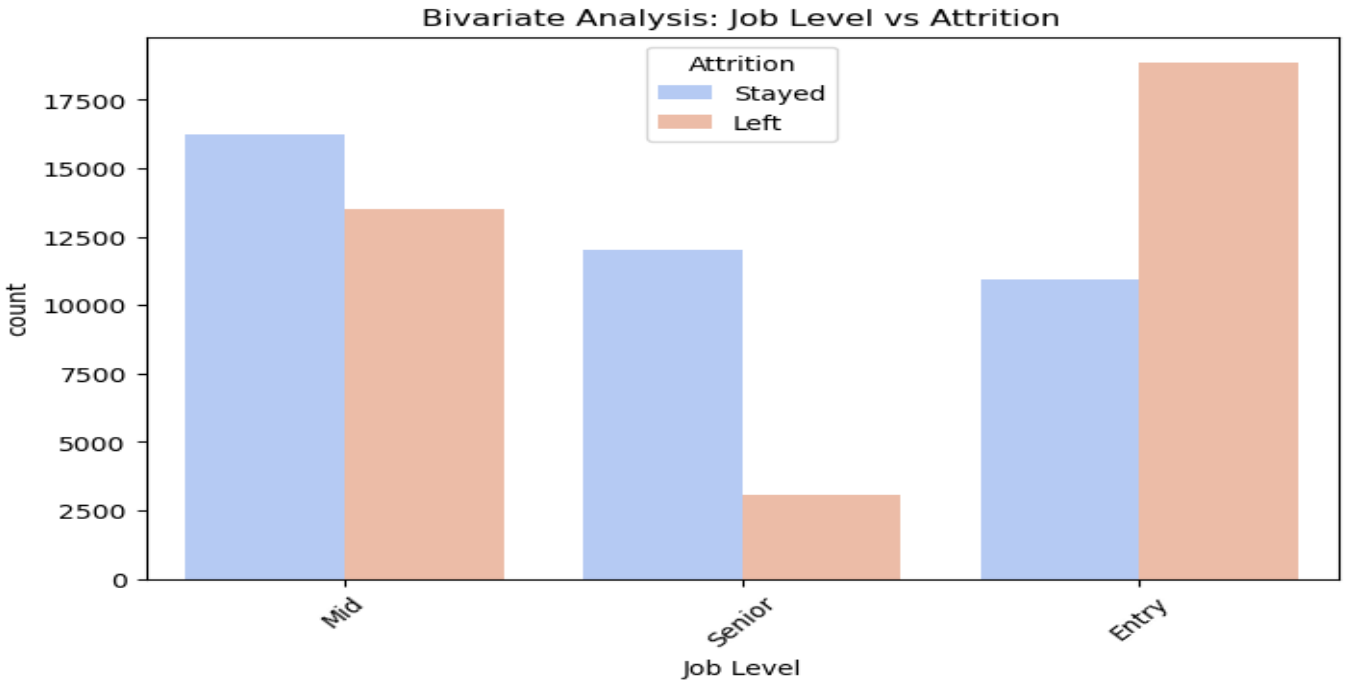
Employees with average performance might display balanced attrition.

4.




Higher attrition among single employees whereas married employees show greater stability resulting in lower attrition rates.

5.




Higher attrition at lower job levels – employees at entry level positions show higher attrition rates, indicating that employees tend to leave more frequently.

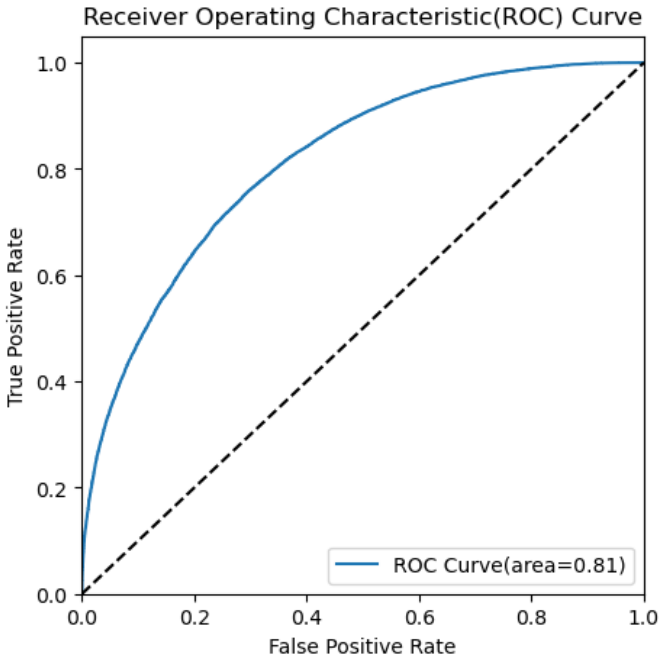
D. Following are the insights of the logistic regression model summary on X_train:

- Pseudo R-Squared (0.2872) – suggests the model explains around 28.7% of the variation in attrition.
- Significant Variables (P < 0.05 for all predictors) – All variables have a strong statistical significance, meaning they contribute meaningfully to the model.
- Log – Likelihood (-27,286) – A lower log-likelihood indicates that the model fits the data reasonably well.

E. Following are the key insights of the result of the Variance Inflation Factor (VIF) output:

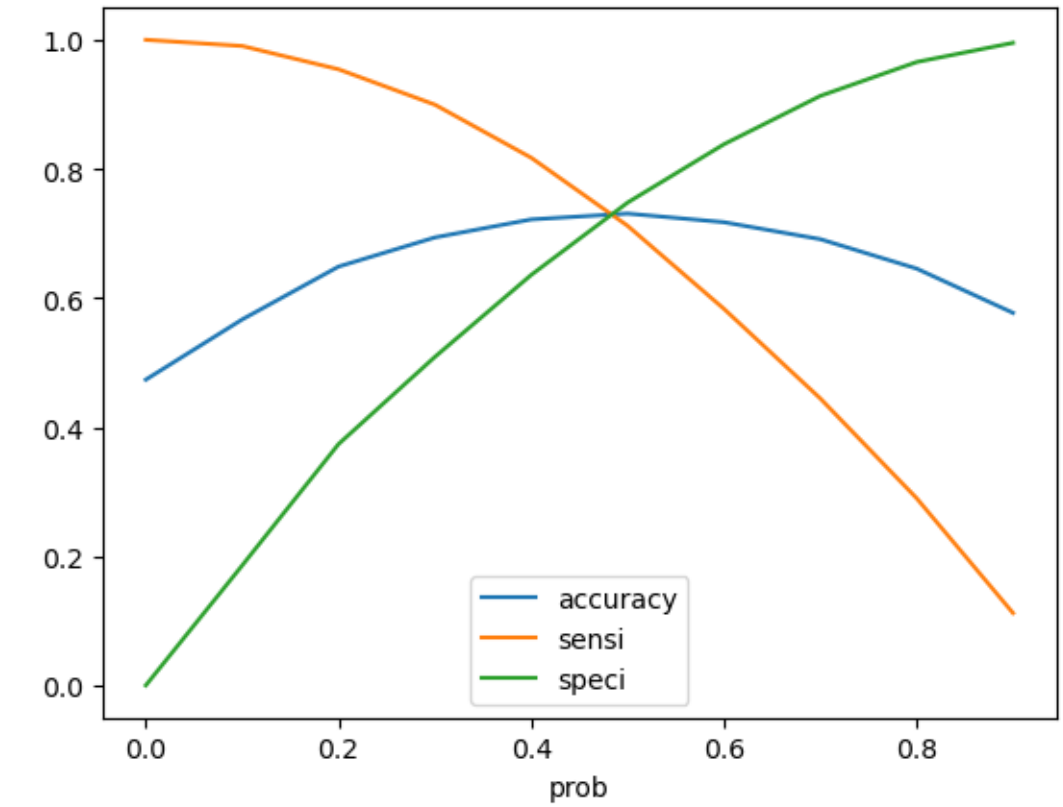
- Low VIF values for most features – Nearly all variables have VIF ~1, suggesting minimal multicollinearity.
- "Job Level_Mid" & "Job Level_Senior" have VIF values above 1.2, indicating a mild correlation but still within acceptable limits.
- VIF of 12.3 for the constant term is expected and does not indicate problematic multicollinearity.

F. ROC Curve:




**Observation:** Higher Area Under the Curve – indicates better model performance, value closer to 1.0 shows excellent classification ability.

This ROC Curve indicates more accurate test.




The above graph shows the optimal cutoff for the final predictions.

G. Conclusion:

## Accuracy (59.6%) – The model correctly predicts attrition in about 60% of cases, which is moderate but not highly reliable.

## Sensitivity (83.37%) – The model effectively detects employees who will leave, meaning it captures attrition cases very well.

## Specificity (37.91%) – The model struggles to correctly classify employees who will stay, which means false positives are quite high.

## Precision (55.09%) – When the model predicts attrition, it's correct about 55% of the time.

## Recall (83.37%) – The model rarely misses attrition cases, making it good at identifying actual leavers.