Predicting Employee Retention

Using Logistic Regression

By
Nagendran T
Naresh Ainapur
Moola Saikiran

Predicting Employee Retention

Objective

To develop a Logistic Regression model to analyse and predict binary outcomes based on the input data

Business Objective

Aim is to provide the HR department with actionable insights to strengthen retention strategies, create a supportive work environment, and increase the overall stability and satisfaction of the workforce.

Assignment Tasks

1. Data Understanding
2. Data Cleaning
3. Train Validation Split
4. EDA on training data
5. Feature Engineering
6. Model Building
7. Prediction and Model Evaluation

Data Cleaning

- **Missing values Columns:** These below columns has missing values

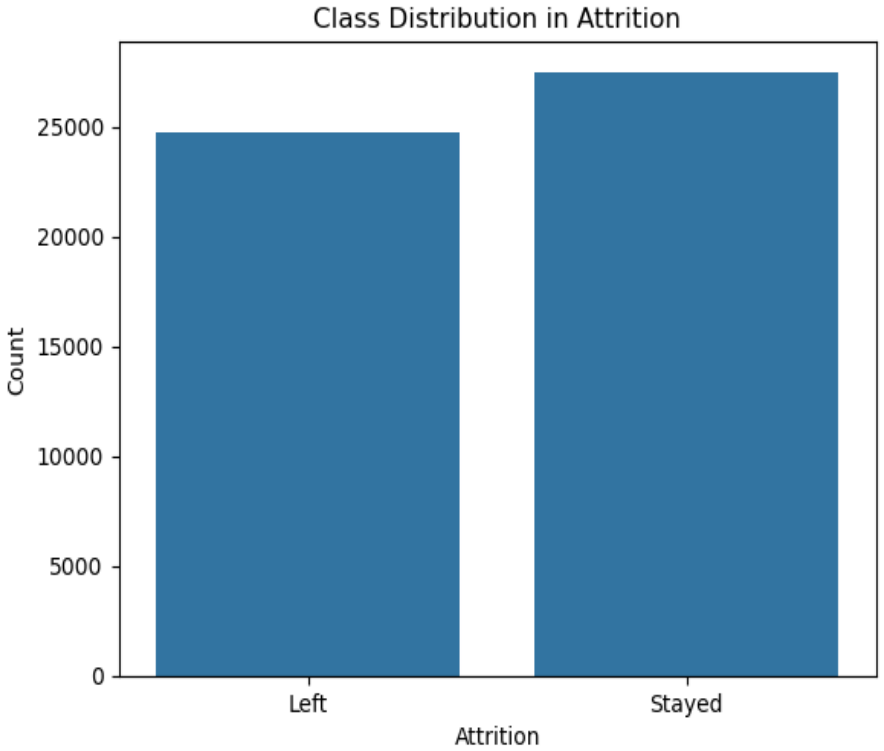
    1. Distance from Home - 2.56%

    2.Company Tenure (In Months)- 3.234151

- Handled missing values with median() method

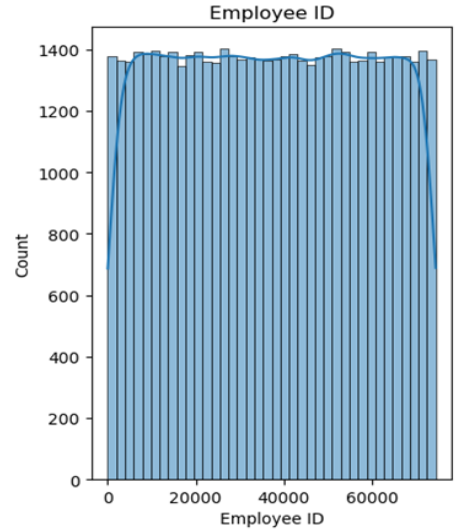
Class Imbalance

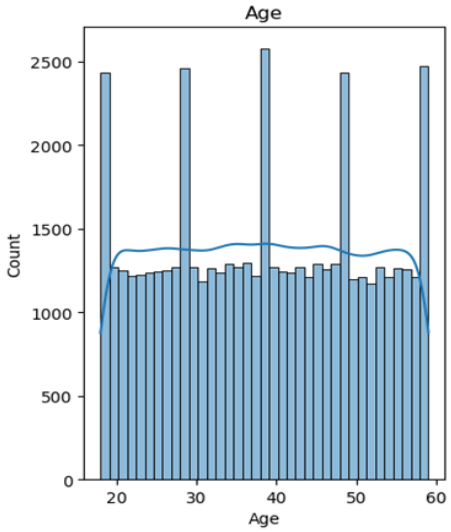
- The number of employees who Stayed is slightly higher than those who Left.
- The distribution is fairly balanced, no large difference between left and stayed.

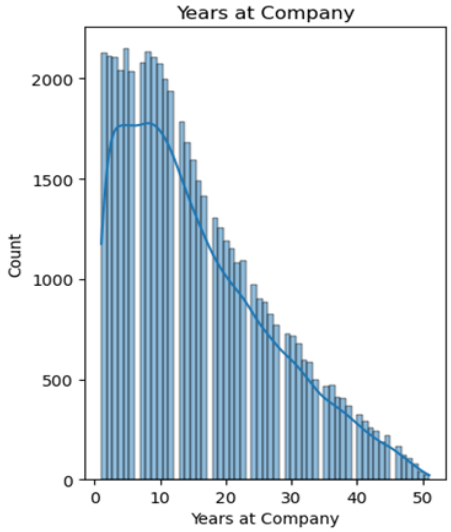


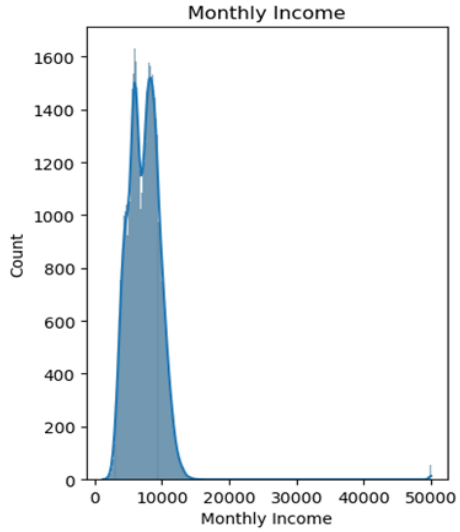
Univariate Analysis

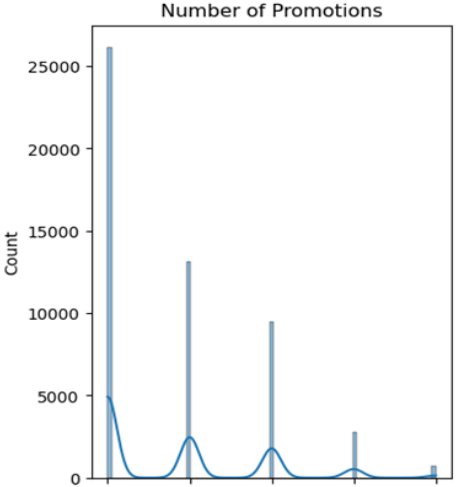
- Performed univariate analysis for all numerical columns
- Numerical Columns are below as given
  'Employee ID',
  'Age',
  'Years at Company',
  'Monthly Income',
  'Number of Promotions',
  'Distance from Home',
  'Number of Dependents',
  'Company Tenure (In Months)'

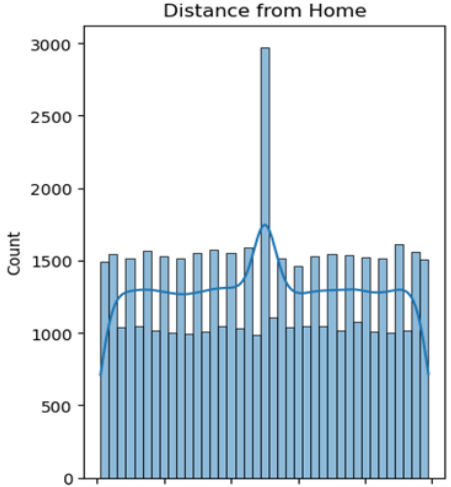

Employee ID
Count
Employee ID

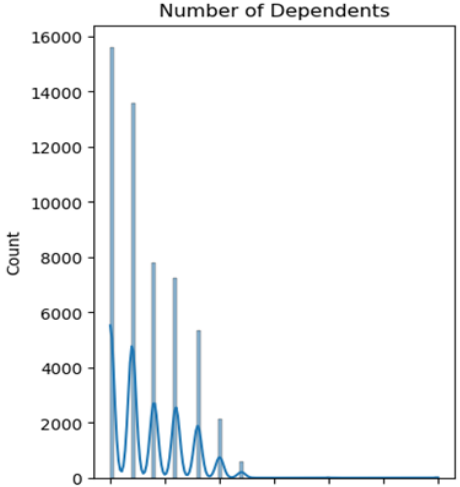

Age
Count
Age

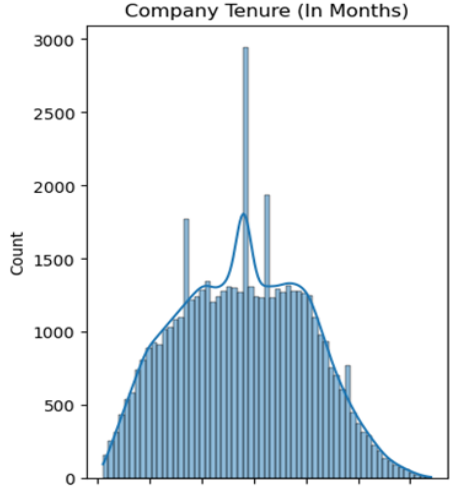

Years at Company
Count
Years at Company


Monthly Income
Count
Monthly Income


Number of Promotions
Count


Distance from Home
Count

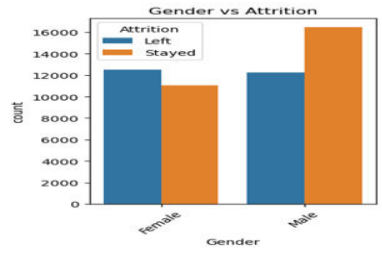

Number of Dependents
Count

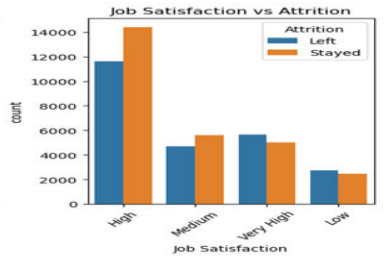

Company Tenure (In Months)
Count

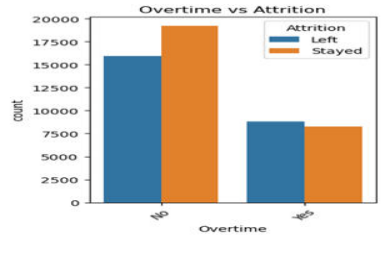
Key Insights from Univariate Analysis

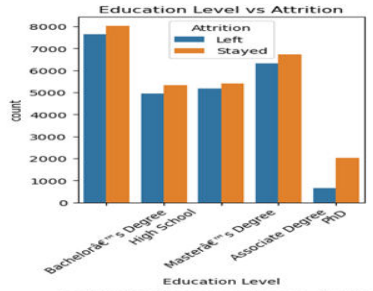
- Employees who has age 39 are more compared to other age groups.
- Employees who has 0-15 years at company is high.
- Average monthly income of all employees is below 20000 per month.
- Employees who have no promotions are more than employees who have 1 or 2 promotions. Once and twice promoted employees are almost same than more than twice promoted.
- The average distance between the employee's home and workplace, in miles is between 40-60.
- The number of dependents on employees is decreasing from 0 to 6. No dependents on employee is high.
- Company tenure(in months) of employees is high between range of 40-60.

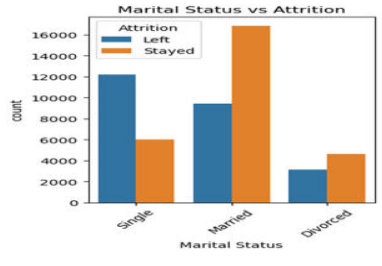
Bivariate Analysis

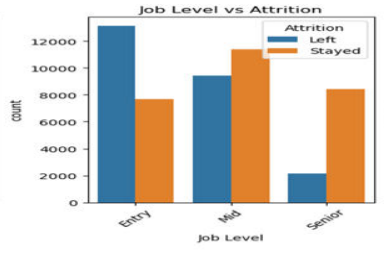


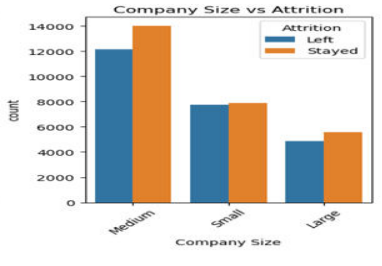


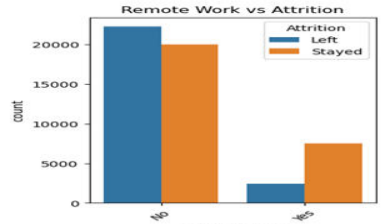


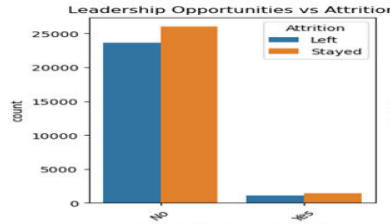


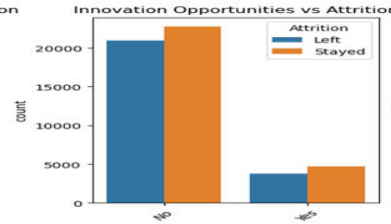


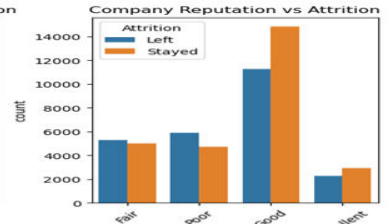


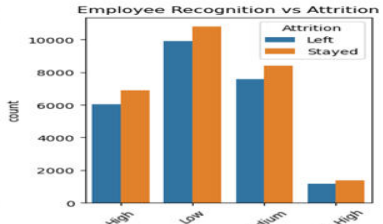












Key Insights of Bivariate Analysis

Demographics and Work Conditions

1. **Gender vs Attrition**: Slightly more males left than females, but the difference is minimal.
2. **Marital Status vs Attrition**: Single employees are more likely to leave compared to married or divorced employees.
3. **Education Level vs Attrition**: Attrition is fairly even across education levels, with slightly higher attrition in those with a Bachelor's or Master's degree.
4. **Company Size vs Attrition**: Medium-sized companies have higher attrition compared to small or large companies.

Key Insights of Bivariate Analysis

5. **Job Role vs Attrition:**

- Sales Executives and Laboratory Technicians show higher attrition rates.
- R&D and Human Resources have lower attrition.

6. **Job Level vs Attrition:**

- Lower job levels (1 & 2) have much higher attrition than higher levels.

7. **Performance Rating vs Attrition:**

- Employees with average ratings show the highest attrition; very few top-rated employees left.

Key Insights of Bivariate Analysis

Work-Life Balance and Satisfaction

8. Work-Life Balance vs Attrition:

- Employees with poor or average work-life balance are more likely to leave.
- Those rating work-life balance as excellent tend to stay.

9. Job Satisfaction vs Attrition:

- Employees with low job satisfaction are much more likely to leave.

10. Overtime vs Attrition:

- Employees working overtime show significantly higher attrition.

Key Insights of Bivariate Analysis

- **Remote and Career Opportunities**

11. **Remote Work vs Attrition:**

- Those with remote work options are less likely to leave.

12. **Leadership Opportunities vs Attrition:**

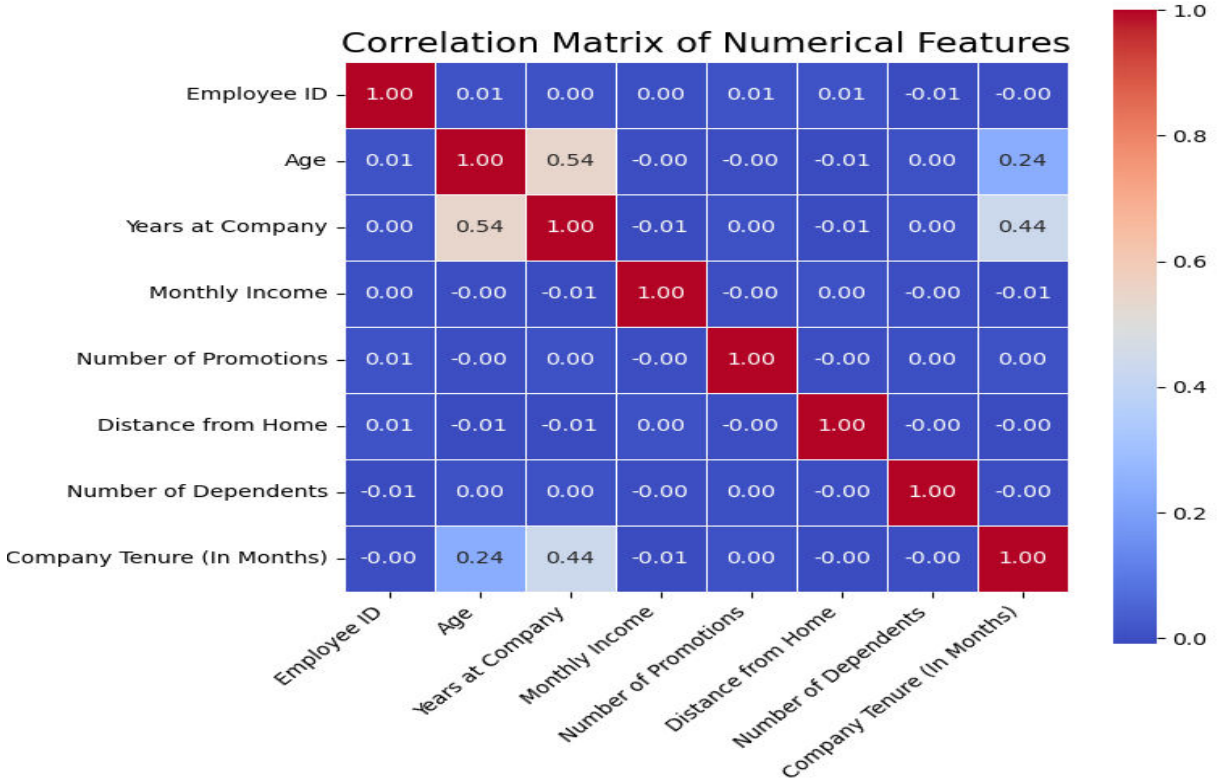
- Lack of leadership opportunities correlates with higher attrition.

13. **Innovation Opportunities vs Attrition:**

- A lack of innovation opportunities leads to higher attrition.

Correlation Analysis

- The features are generally not highly correlated.
- Only a few variables like Age, Years at Company, and Tenure are somewhat interrelated, which makes sense logically and may help explain behavior like retention.



Train-Validation Split

- **Feature Variables in X :** 'Age', 'Gender', 'Years at Company', 'Job Role', 'Monthly Income', 'Work-Life Balance', 'Job Satisfaction', 'Performance Rating', 'Number of Promotions', 'Distance from Home', 'Education Level', 'Marital Status', 'Number of Dependents', 'Job Level', 'Company Size', 'Company Tenure (In Months)', 'Remote Work', 'Leadership Opportunities' 'Innovation Opportunities', 'Company Reputation', 'Employee Recognition'

- **Target Variable in Y:** 'Attrition'

- Splitted the data into 70% train data and 30% validation data

Feature Engineering and Feature scaling

- Created dummy variables for categorical columns in both training and validation datasets.

- Applied feature scaling to the numeric columns for consistent scaling to both X_train and X_test.

Test-Validation Split

- **Feature Variables in X :** 'Age', 'Gender', 'Years at Company', 'Job Role', 'Monthly Income', 'Work-Life Balance', 'Job Satisfaction', 'Performance Rating', 'Number of Promotions', 'Distance from Home', 'Education Level', 'Marital Status', 'Number of Dependents', 'Job Level', 'Company Size', 'Company Tenure (In Months)', 'Remote Work', 'Leadership Opportunities' 'Innovation Opportunities', 'Company Reputation', 'Employee Recognition'

- **Target Variable in Y:** 'Attrition'

- Splitted the data into 70% train data and 30% validation data

Feature selection

- By using Recursive Feature Elimination (RFE), selected the most influential features for building the model. Those are as follows
  'Remote Work', 'Gender_Male', 'Work-Life Balance_Fair', 'Work-Life Balance_Good', 'Work-Life Balance_Poor', 'Job Satisfaction_Low', 'Job Satisfaction_Very High', 'Performance Rating_Below Average', 'Performance Rating_Low', 'Education Level_PhD', 'Marital Status_Single', 'Job Level_Mid', 'Job Level_Senior', 'Company Reputation_Fair', 'Company Reputation_Poor'.

Building Logistic Regression Model

- By using Statsmodel built logistic regression, this enables to evaluate statistical metrics like p-values and VIFs, which are important for identifying multicollinearity.
- All the p-values are less than 0.05
- All VIF values of feature variables are less than 5
- There’s no requirement of dropping variables

Building Logistic Regression Model

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.3250	0.034	9.572	0.000	0.258	0.392
Remote Work	1.7039	0.030	56.293	0.000	1.645	1.763
Gender_Male	0.5899	0.022	27.331	0.000	0.548	0.632
Work-Life Balance_Fair	-1.2740	0.032	-39.218	0.000	-1.338	-1.210
Work-Life Balance_Good	-0.3027	0.031	-9.823	0.000	-0.363	-0.242
Work-Life Balance_Poor	-1.4303	0.039	-36.627	0.000	-1.507	-1.354
Job Satisfaction_Low	-0.4602	0.036	-12.695	0.000	-0.531	-0.389
Job Satisfaction_Very High	-0.4738	0.027	-17.656	0.000	-0.526	-0.421
Performance Rating_Below Average	-0.3192	0.030	-10.628	0.000	-0.378	-0.260
Performance Rating_Low	-0.5712	0.049	-11.591	0.000	-0.668	-0.475
Education Level_PhD	1.5157	0.054	27.903	0.000	1.409	1.622
Marital Status_Single	-1.7007	0.024	-71.006	0.000	-1.748	-1.654
Job Level_Mid	0.9514	0.024	40.464	0.000	0.905	0.998
Job Level_Senior	2.5034	0.034	74.375	0.000	2.437	2.569
Company Reputation_Fair	-0.4868	0.028	-17.559	0.000	-0.541	-0.432
Company Reputation_Poor	-0.7219	0.028	-26.122	0.000	-0.776	-0.668

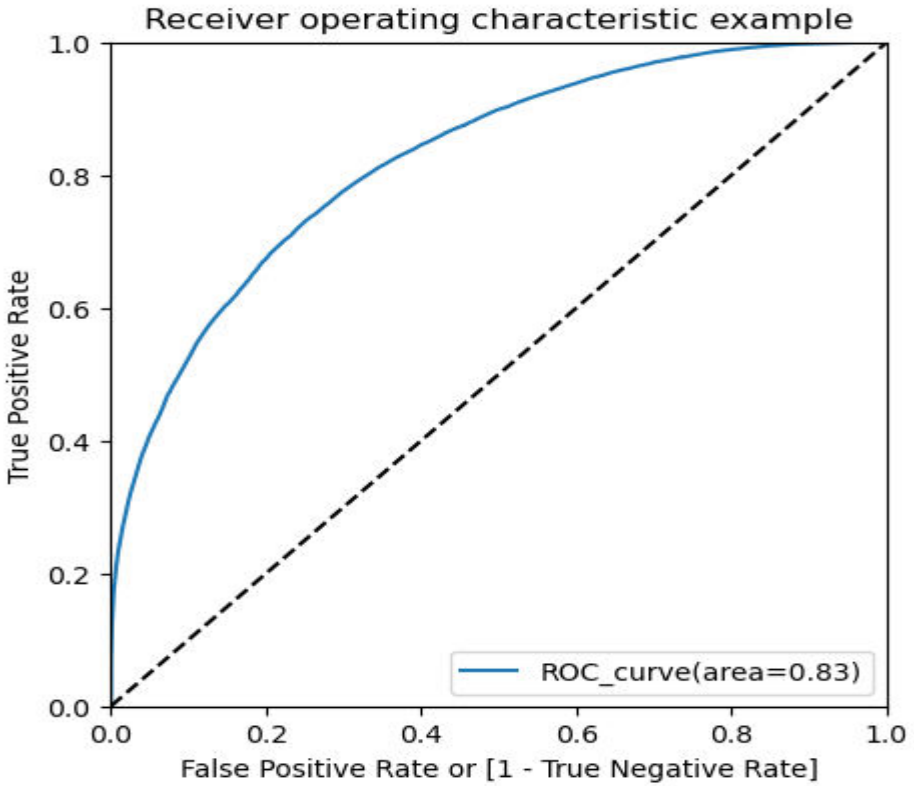
Predictions on training set

- Predicted the output with X_train
- Created a new DataFrame containing the actual stayed flag and the probabilities predicted by the model
- Created a new column 'Predicted' with 1 if
- predicted probabilities are greater than 0.5 else 0
- This picture is dataframe of y_train_pred_df
- Actual and Predicted output are matching

	Act_Attrition	Pred_Attrition	Employee ID	Predicted
60704	1	0.966128	60704	1
16163	1	0.903584	16163	1
25709	0	0.303925	25709	0
4354	1	0.559003	4354	1
49862	1	0.646564	49862	1

Optimal Cutoff

- AUC (Area Under Curve): 0.83
- 83% chance the model randomly chosen true positives higher than a false positive one.



Evaluation of Performance of Model

Below metrics are based on the predictions made on the training set

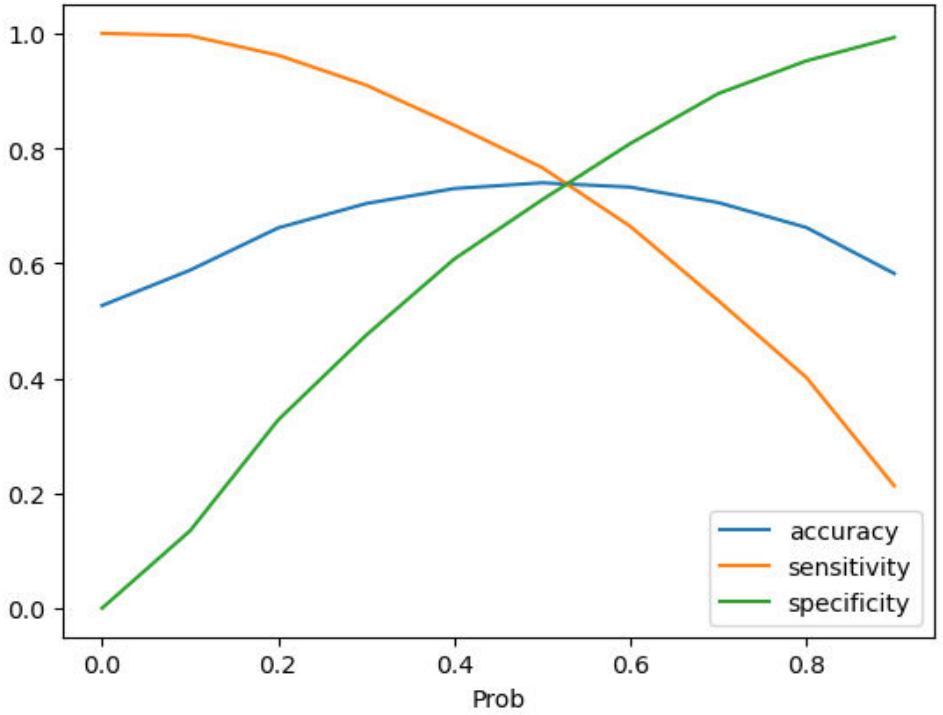
- Confusion Matrix : array([[17572, 7160],
  [6421, 21074]]

Accuracy	0.73
Sensitivity	0.76
Specificity	0.71
Precision	0.74
Recall	0.76

Sensitivity and Specificity TradeOff

This plot shows how accuracy, sensitivity, and specificity vary with different probability thresholds.

- Accuracy : Peaks around 0.5 threshold and then drops off
- Sensitivity : High at low thresholds and decreases as threshold increases.
- Specificity: Low at low thresholds (many false positives), and increases with higher thresholds
- Optimal threshold is where accuracy sensitivity and specificity intersect.



Evaluation of Performance of Model

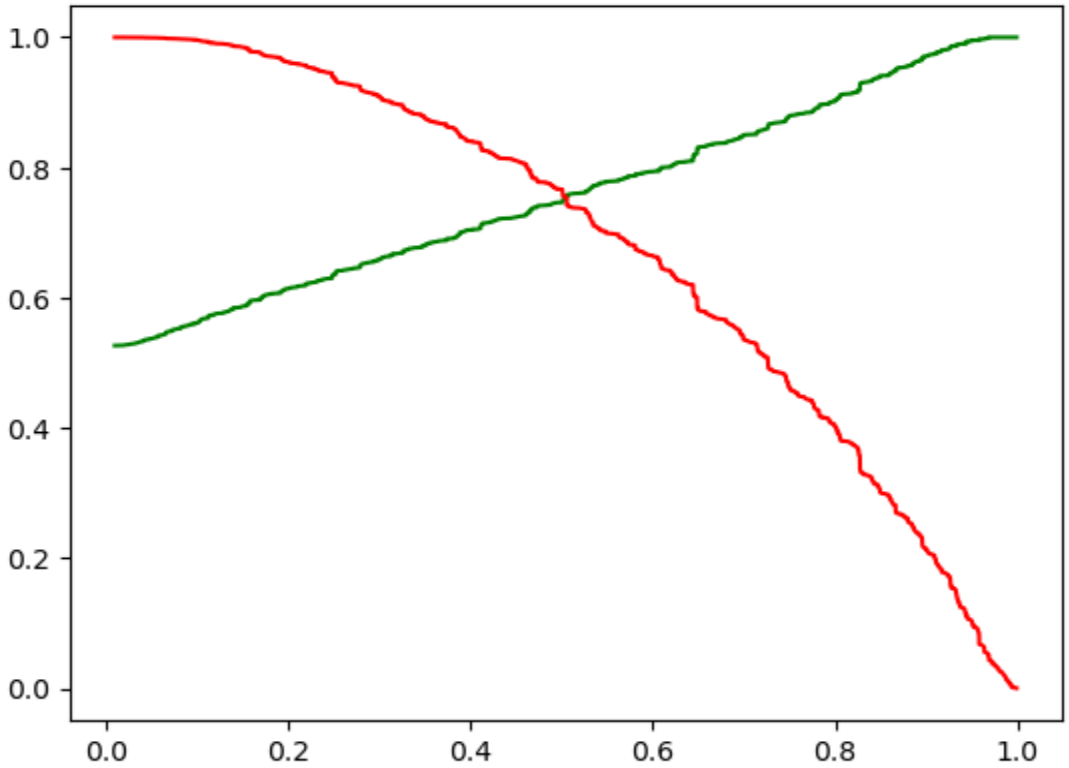
Created a column for final prediction based on the optimal cutoff

- Confusion Matrix : array([[17572, 7160],
  [6421, 21074]]

Accuracy	0.73
Sensitivity	0.76
Specificity	0.71
Precision	0.74
Recall	0.76

Precision and Recall Tradeoff

- The point where the two curves intersect is often considered a balanced threshold
- The threshold is 0.5



Prediction and Model Evaluation

- Predictions on the validation set using the optimal cutoff and stored in 'final_prediction' column.
- Actual and Predicted are matching.

	actual	predicted_probability	final_prediction
0	1	0.986814	1
1	1	0.725859	1
2	0	0.175269	0
3	0	0.384308	0
4	0	0.058327	0

Evaluation of Performance of Model

- Confusion Matrix : array([[7603, 3084],
  [2793, 8903]]

Accuracy	0.73
Sensitivity	0.76
Specificity	0.71
Precision	0.74
Recall	0.76

Conclusion

- **Factors Increasing Retention (Positive Coefficients):**

Job Level_Senior(Coef= 2.5034): Senior-level employees have the highest retention — indicates strong loyalty or role satisfaction. Remote Work(Coef= 1.7039): Remote workers are more likely to stay — indicates strong engagement remotely. Education Level_PhD(Coef= 1.5157): PhDs are more likely to stay — suggests high satisfaction or alignment with role. Job Level_Mid(Coef= 0.9514): Mid-level employees are more likely to stay. Gender_Male(Coef= 0.5899): Males are more likely to stay than females.

- **Factors Increasing Attrition (Negative Coefficients):**

Company Reputation_Poor(Coef= -0.7219): Poor company reputation increases attrition. Marital Status_Single(Coef= -1.7007): Single employees are more likely to leave. Work-Life Balance_Poor(Coef= -1.4303): Poor work-life balance strongly drives attrition. Performance Rating_Low(Coef= -0.5712): Lower performers are more likely to leave — could reflect management action or disengagement. Job Satisfaction_Low(Coef= -0.4602): Lower satisfaction leads to higher attrition.