Case Study

Predicting Employee Retention Using Logistic Regression

1.Problem Statement

Employee turnover poses significant challenges for organizations, leading to increased costs and loss of talent. Identifying the key factors that influence an employee's decision to stay or leave is crucial for improving retention. This project aims to analyze employee data and develop a **Logistic Regression model** to predict attrition based on variables such as age, job role, satisfaction, work-life balance, income, and company-related factors.

2. Methodology

2.1 Dataset Overview

- Total records: 74,610
- Total features: 24
- Target variable: Attrition ("Stayed" or "Left")
- Features include a mix of categorical and numerical data
- Class distribution is relatively balanced, supporting unbiased model training

2.2 Data Understanding:

The dataset comprises 24 features, with the target variable being *Attrition*, indicating whether an employee has “Stayed” or “Left” the organization.

2.3 Data Cleaning:

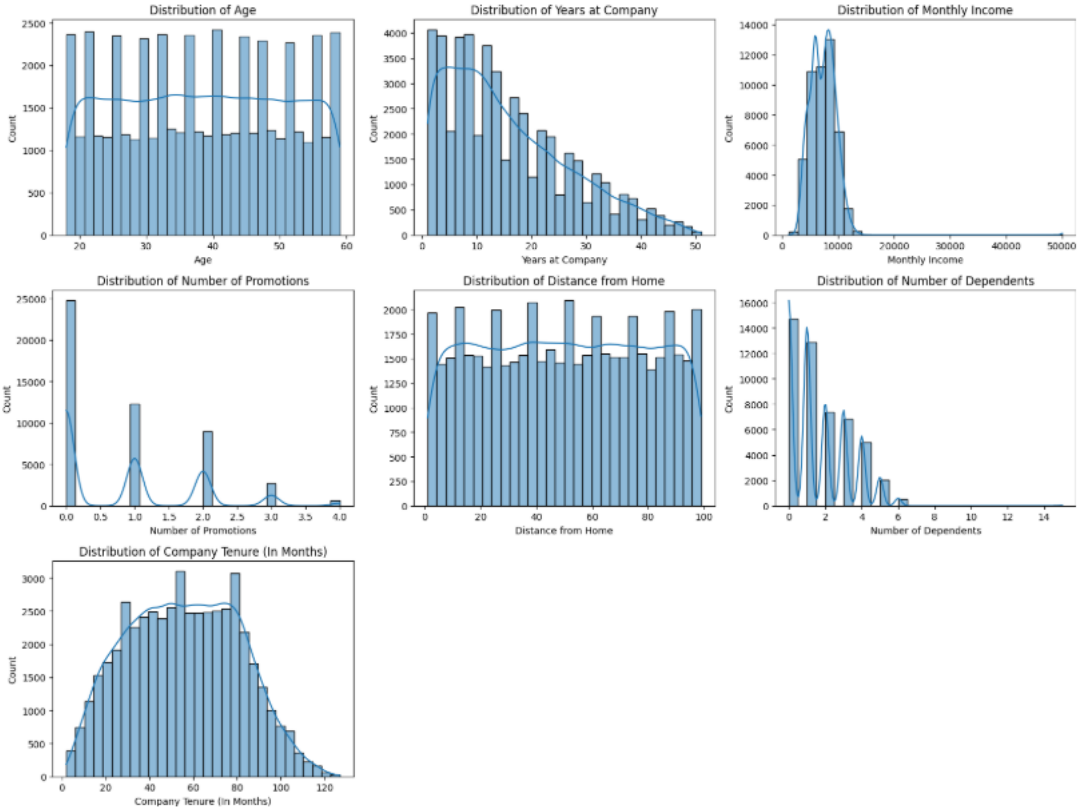
The *Distance from Home* column had 1,912 missing values (approximately 2.56% of the data), while *Company Tenure (In Months)* had 2,413 missing entries (around 3.23%). To handle these missing values and reduce bias, imputation methods such as mean or median replacement were applied. Categorical variables like *Gender* and *Job Role* were thoroughly reviewed to identify and eliminate any redundancies or duplicate records, ensuring the dataset was clean and ready for model development.

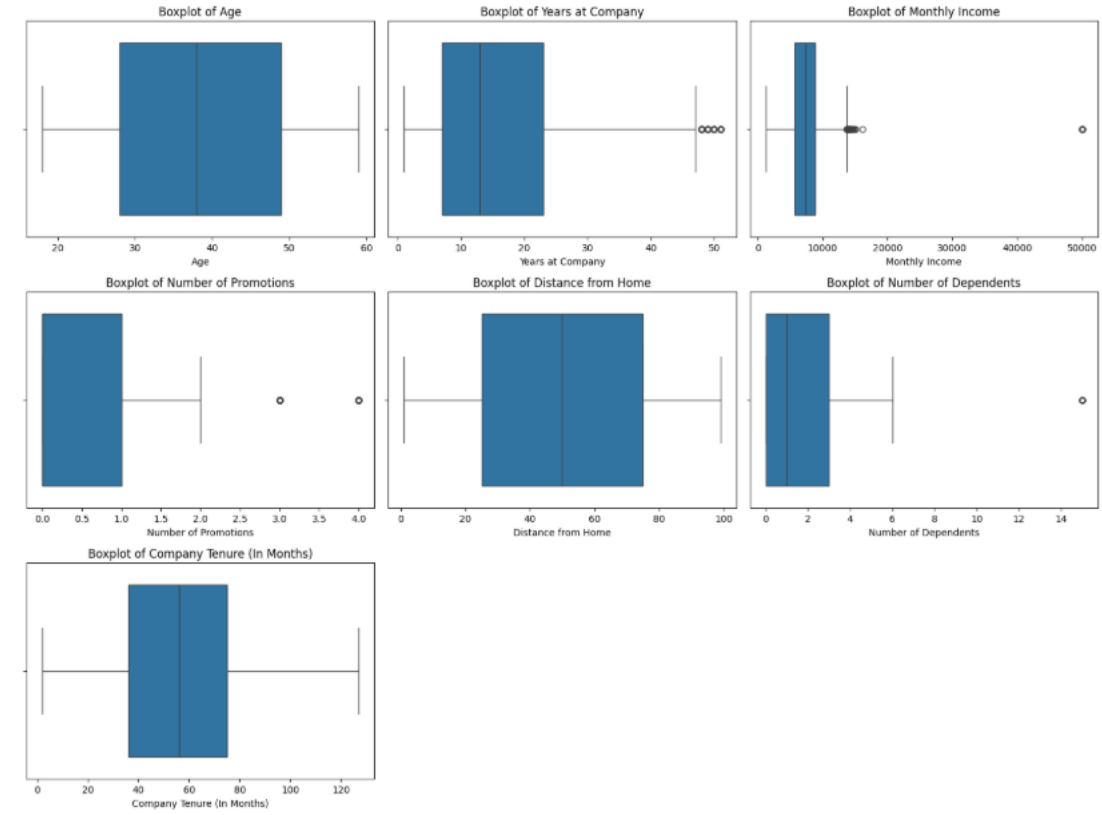
2.4 Train-Validation Split:

The dataset was divided into training and validation sets using a 70:30 split, enabling effective model training and performance evaluation.

2.5 EDA on Training Data:

2.5.1 Histogram for each numerical column



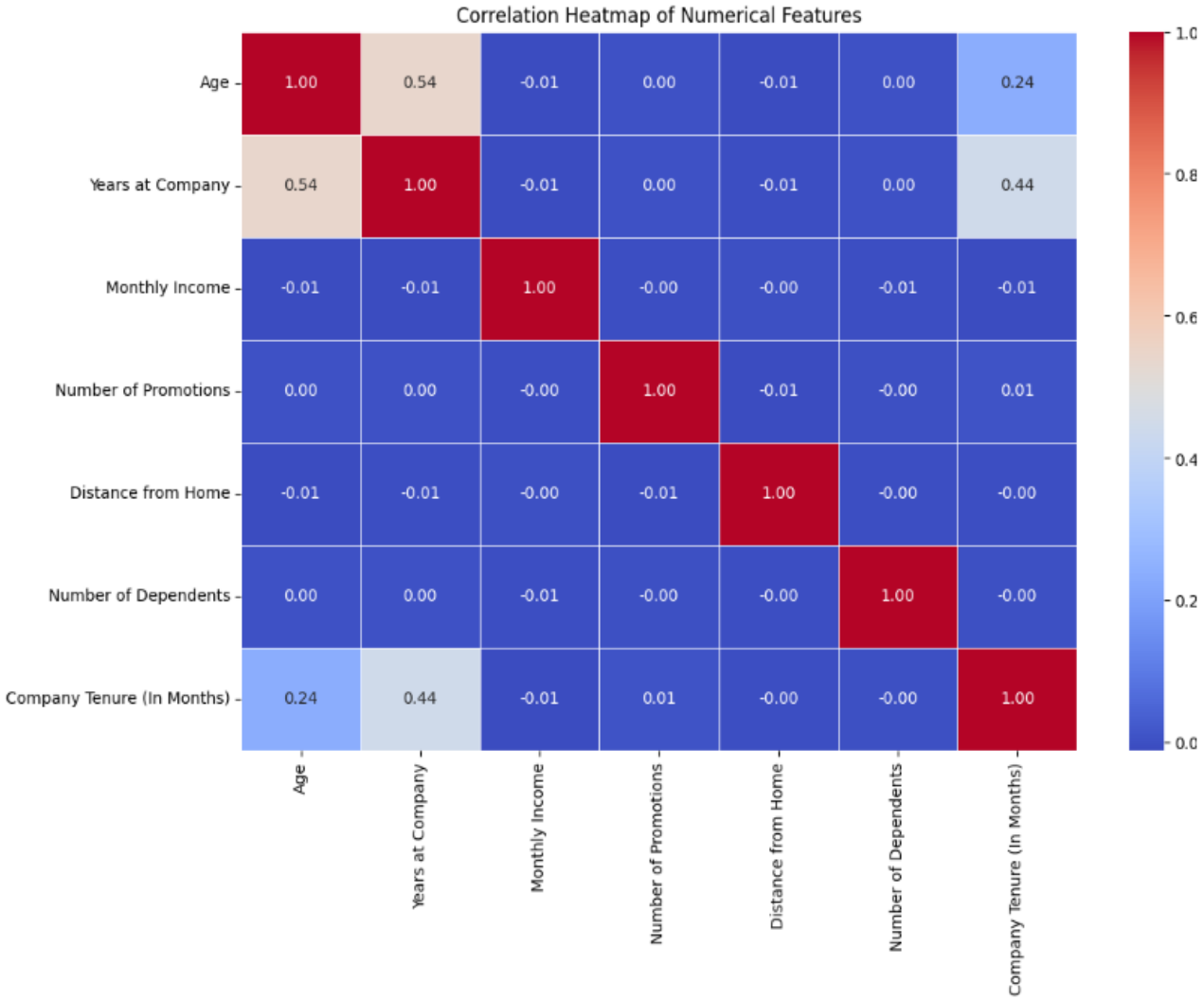


Insights:

- The age distribution appears fairly balanced among mid-career employees, whereas *Years at Company* and *Monthly Income* show a right-skewed pattern, suggesting a higher number of employees with shorter tenures and lower salaries. Additionally, most employees have received few promotions, live a moderate distance from the workplace, and have between zero to two dependents.

- Box plots support the histogram findings by clearly displaying the central tendencies and variability in the data. *Tenure* and *income* remain notably right-skewed, while promotions and dependents are clustered at lower values. Outliers are observed in several variables—including *Years at Company*, *Monthly Income*, *Distance from Home*, and *Number of Dependents*—indicating the presence of employees with significantly different profiles.

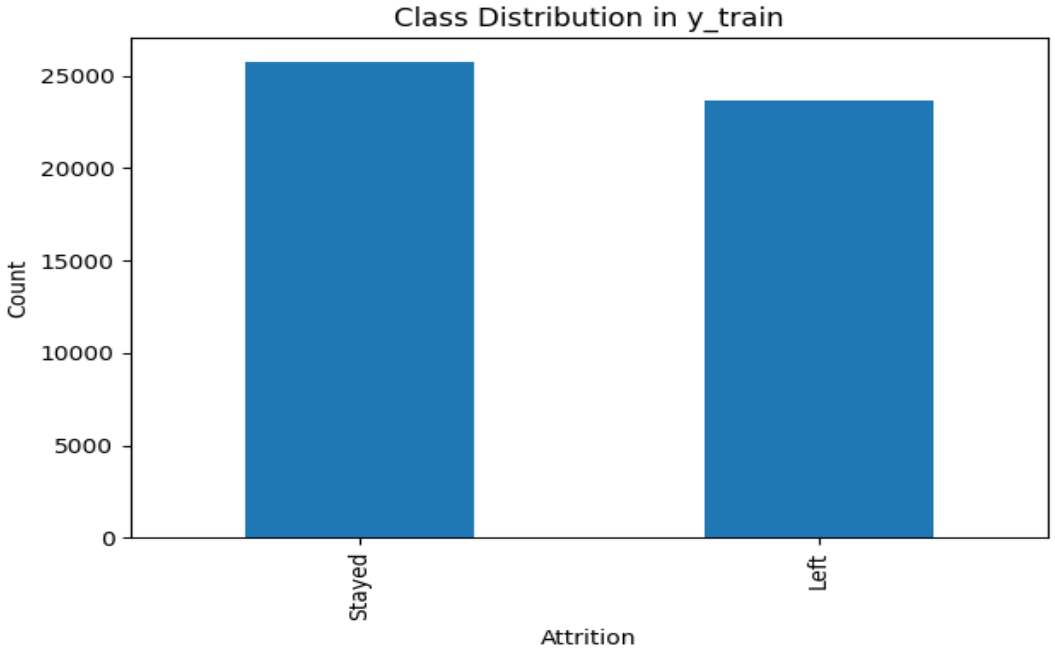
2.5.2 Heatmap of the correlation matrix



Insights:

A moderate positive correlation (0.44) is observed between *Years at Company* and *Company Tenure (in Months)*, which is expected since both represent similar information in different units. Additionally, *Age* and *Years at Company* show a weaker positive correlation (0.54), indicating that older employees are generally more likely to have longer tenures with the organization.

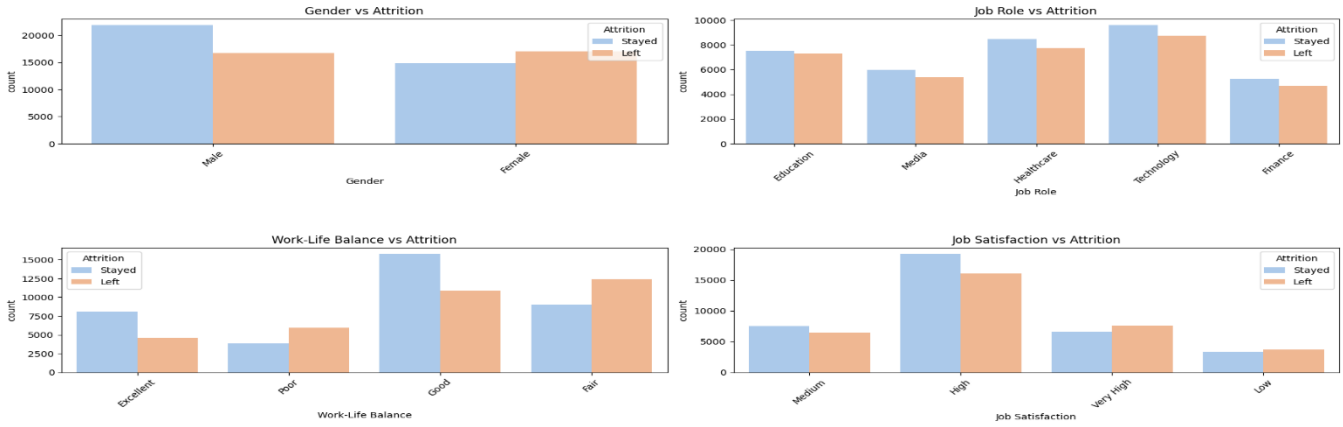
2.5.3 Bar graph to check class balance

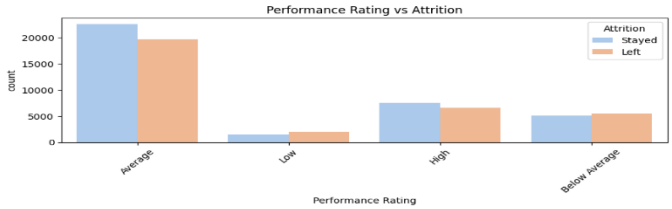


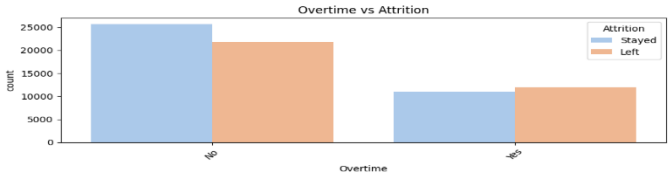
Insights:

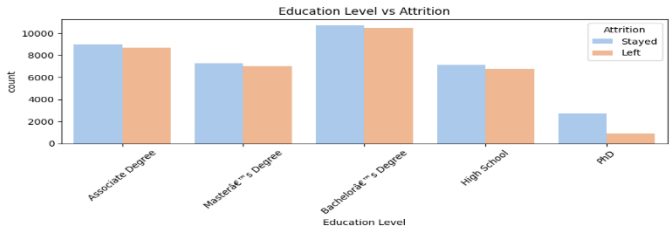
The bar chart shows a relatively balanced distribution of the target variable in the training set, with a slightly higher count for "Stayed" compared to "Left" . This near-equal representation of both classes is generally good for training a classification model as it helps prevent bias towards the majority class.

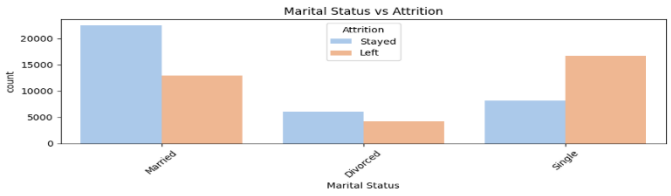
3.5.4 Bivariate Analysis on training data between all the categorical columns and target variable

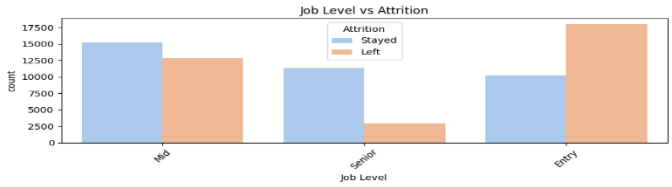


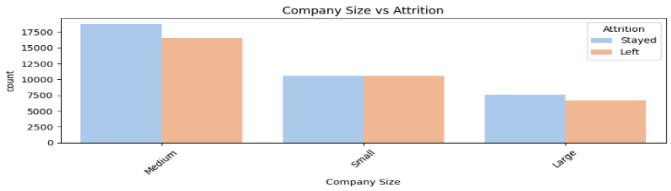


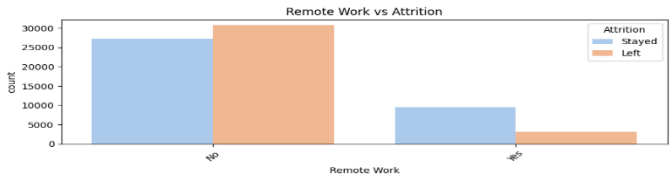


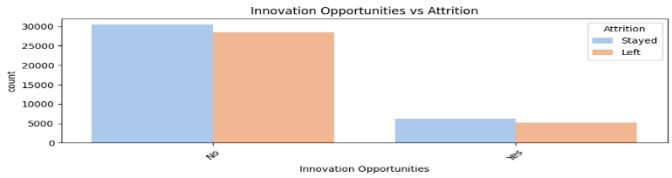


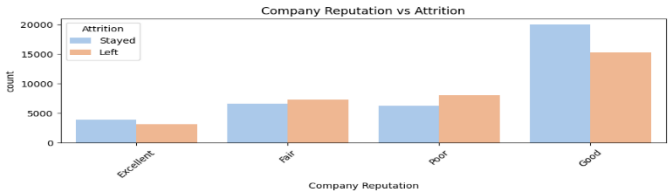












Insights:

The charts illustrate how attrition rates differ across various categories within each employee attribute. For example, certain genders, job roles, or work-life balance levels may exhibit higher or lower tendencies for employees to leave the company.

3. Feature Engineering:

- At the outset, categorical, independent, and dependent columns were identified to optimize the preprocessing process. Dummy encoding was applied to the categorical variables in both the training and validation datasets to maintain consistency. After encoding, the original categorical columns were removed from both datasets to eliminate redundancy. Additionally, other

unnecessary or irrelevant columns were dropped to improve data quality and minimize potential noise during model training.

- Feature scaling was applied to the numeric columns to standardize their range, ensuring consistent scaling across the dataset.

4. Model Building:

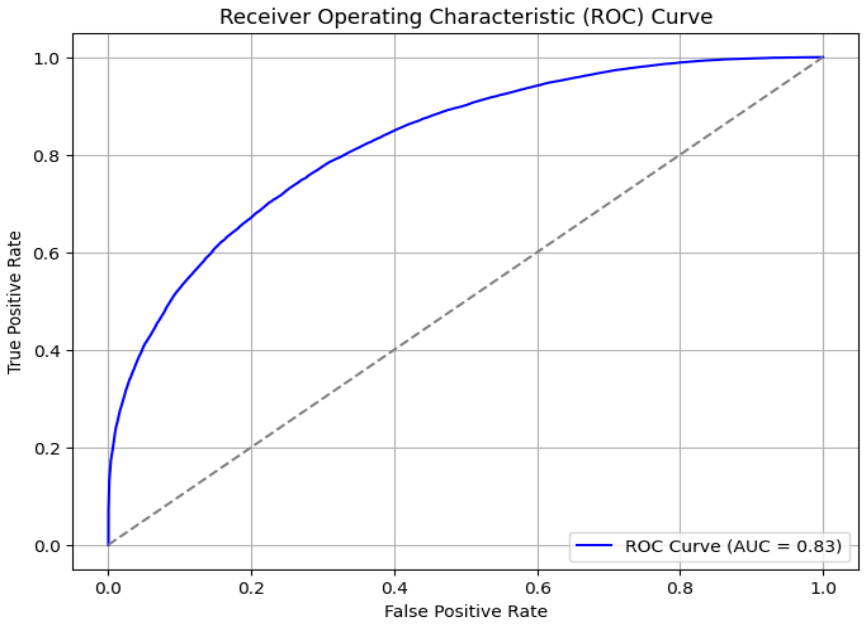
- Feature selection was carried out using the Recursive Feature Elimination (RFE) technique to identify the most important variables for the model. The chosen features were then displayed for review.
- A Logistic Regression model was built using the selected features and trained on the training data.
- To evaluate multicollinearity, the Variance Inflation Factor (VIF) was computed using the *variance_inflation_factor* function from the *statsmodels.stats.outliers_influence* module. Features with VIF values exceeding acceptable thresholds were closely examined.
- If any feature had a high VIF and did not meet the required p-value significance, it was removed from the dataset. The model was then retrained to ensure optimal feature selection.
- After finalizing the model, predictions were made on the validation set.
- The model's performance was evaluated using metrics from the scikit-learn library, achieving an overall accuracy of 73.94%.
- The Confusion Matrix for the model was:

[[17052 6593]

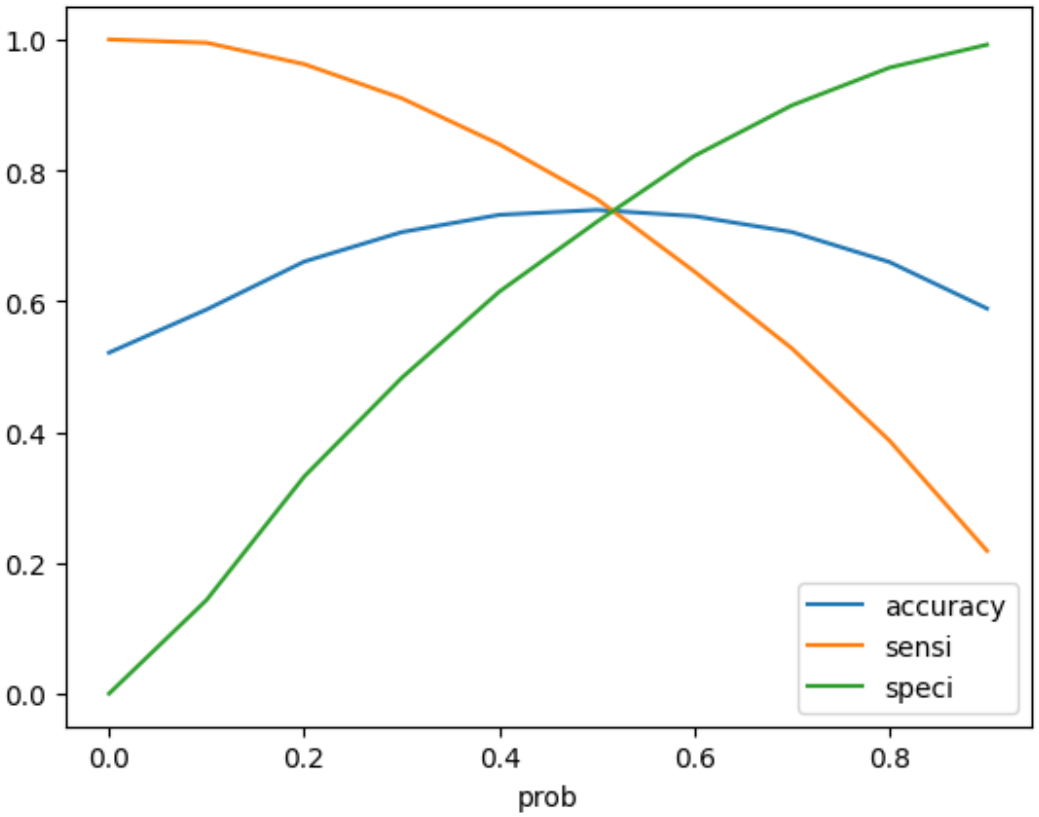
[6278 19467]]

Additional performance metrics were calculated:

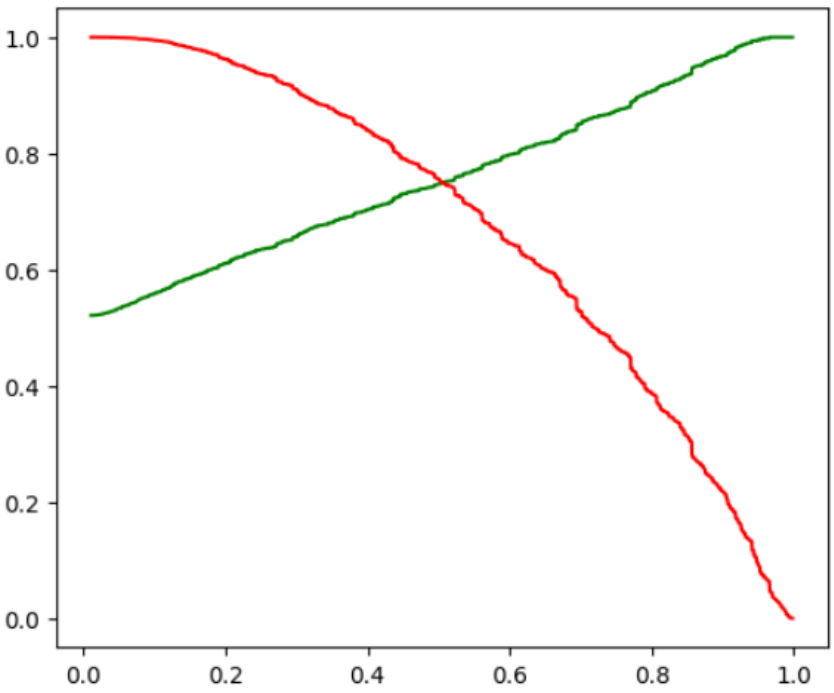
- Sensitivity: 0.7561
- Specificity: 0.7212
- Precision: 0.747
- Recall: 0.7561



- The Receiver Operating Characteristic (ROC) curve was generated to identify the optimal probability cutoff value for classifying outcomes.



- Additional plots were created to visualize the Accuracy, Sensitivity, Specificity, Precision, and Recall at various probability cutoff thresholds, providing a more comprehensive evaluation.



5. Prediction and Model Evaluation:

- The final step involved testing the trained model on the validation dataset to assess its performance in real-world scenarios.
- Relevant features, as selected during the feature selection phase, were extracted from the validation dataset to maintain consistency with the training process.
- Predictions were made on the validation set using the previously determined optimal cutoff value.
- The model achieved an accuracy of 73.55% on the validation data.
- The Confusion Matrix for the validation set was:

  [[7829 2313]
  [3286 7740]]
- Additional performance metrics were calculated to evaluate the model:
  - Sensitivity : 0.7019
  - Specificity: 0.7719
  - Precision: 0.7699
  - Recall: 0.702

6. Assumptions Made

- The dataset provides a balanced representation of the company's employees.

- Categorical values were consistent and appropriately encoded.
- Logistic Regression was selected for its interpretability; although more complex models might offer improved results, they were not explored due to project scope constraints.

7. Conclusion

In this project, we developed a Logistic Regression model to predict employee attrition based on various employee attributes such as demographics, job satisfaction, and work-life balance. After preprocessing the data, handling missing values, and encoding categorical variables, we performed feature selection using Recursive Feature Elimination (RFE) and addressed multicollinearity using the Variance Inflation Factor (VIF). The model was trained and evaluated using accuracy, precision, recall, and specificity, achieving an accuracy of 73.94% on the training set and 73.55% on the validation set. This project highlights the importance of key factors such as work-life balance and job satisfaction in predicting employee attrition, providing valuable insights for organizations to improve retention strategies.